UNIVERSITE DE DROIT ET DES SCIENCES D'AIX-MARSEILLE III
Faculté des sciences et Techniques de Saint Jérôme

THESE

Pour obtenir le grade de

Docteur de l'université Aix-Marseille III

Discipline : Sciences de l'Information et de la Communication

par

Hélène ZIEGELBAUM

Le 28 mai 1998

Nouvelles approches dans la recherche d'innovation en agroalimentaire

Mise au point et valorisation de nouvelles procédures pour mieux connaître la perception des produits alimentaires par les consommateurs

Directeur de thèse

Henri DOU

JURY

Henri DOU, *Professeur Université Aix-Marseille III (Examinateur)*
Jean MOSCAROLA, *Professeur Université de Savoie (Rapporteur)*
Joseph HOSSENLOPP, *Professeur ENSAM (Rapporteur)*
Hervé ROSTAING, *Maître de Conférences Université Aix-Marseille III (Examinateur)*
Michel ROGEAUX, *Ingénieur TEPRAL/DANONE (Examinateur)*

Cette version, antérieure à la soutenance, demeure provisoire et soumise aux éventuelles modifications demandées par le jury.

Ce travail n'aurait sans doute jamais aboutit de cette façon sans la collaboration et l'aide de certaines personnes que je tiens à remercier :

Michel CARNIELO, Directeur Général du TEPRAL et Henri DOU, directeur du CRRM grâce à qui, tous les moyens ont été mis en place,

Jean MOSCAROLA et Joseph HOSSENLOPP pour avoir accepté de valider ce travail en tant que rapporteurs,

Michel ROGEAUX, Responsable du service Sciences du Goût au TEPRAL pour tout ce qu'il a pu m'apprendre et la confiance qu'il m'a accordée,

Hervé ROSTAING, Maître de Conférence au CRRM pour son sérieux dans le suivi de mes travaux malgré notre éloignement,

Les membres du réseau Analyse Sensorielle du Groupe DANONE qui m'ont procuré des données pour évaluer la méthode de traitement,

Toute l'équipe des Sciences du Goût du TEPRAL, pour la qualité de l'ambiance de travail,

Toute l'équipe du CRRM, Luc QUONIAM, Eric GIRAUD, Sandrine ESTATICO et tous les étudiants thésards pour authenticité marseillaise qui me « fendait le cœur » à chaque appel téléphonique et visite,

Enfin, je remercie l'ensemble du personnel du TEPRAL pour leur sympathie et leur accueil dans leur centre de recherche.

RESUME

Le marché de l'agroalimentaire d'aujourd'hui rencontre les difficultés liées à la concurrence et aux exigences des consommateurs. Dans ce contexte, plusieurs solutions s'offrent aux professionnels de l'alimentation qui veulent être compétitifs : la gestion de la qualité des produits existants et le développement de nouveaux produits dans le respect des réglementations en vigueur.
Cette thèse expose une démarche particulière de recherche d'innovation à travers la connaissance de la perception des produits alimentaires par les consommateurs.
Classiquement, la collecte de l'information consommateur est réalisée grâce à des tests pendant lesquels les produits sont dégustés. Nous avons eu l'occasion de remarquer que l'emploi des questions fermées dans ces tests sensoriels n'était pas adapté aux consommateurs. En effet, ces derniers peuvent être influencés par les réponses qui leur sont proposées et l'image du produit transmise par le consommateur ne sera plus en adéquation avec ses réelles perceptions.
Aussi, pour donner aux consommateurs une entière spontanéité, des questions ouvertes ont été introduites. Le consommateur s'exprime librement sur ses perceptions vis à vis du produit. Il utilise son propre langage et associe intuitivement des termes à ses sensations. De cette façon, l'image du produit chez le consommateur est transmise fidèlement.
Cette information est très importante pour les professionnels de l'agroalimentaire. En effet elle permet, d'une part, de mieux communiquer sur ses produits et, d'autre part, de mieux connaître les attentes des consommateurs.
Pour exploiter au mieux cette information riche mais complexe issue des questions ouvertes, nous avons dû mettre au point une méthode de traitement spécifique afin d'obtenir une information homogène et fiable à partir de texte brut.
Cette méthode s'est inspirée des techniques d'analyse en bibliométrie et en lexicométrie. Des programmes informatiques simples ont été développés pour l'automatiser.

Mots Clés *:* Innovation - Agroalimentaire - Consommateur - Analyse sensorielle
Statistiques textuelles - Bibliométrie - Veille produit -

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

INTRODUCTION

Dans un contexte de mondialisation des marchés, les industries agroalimentaires évoluent à l'heure actuelle dans un environnement de guerre concurrentielle. Elles doivent donc s'armer de nouvelles solutions de compétition pour rester dans la course ou pour se positionner parmi les premiers.
Cette thèse propose une nouvelle approche dans la recherche d'innovation en agroalimentaire grâce à la mise au point et à la valorisation de procédures pour mieux connaître la perception des produits alimentaires par les consommateurs.

Le premier chapitre tente de situer le contexte du projet de recherche réalisé au cours de cette thèse en établissant tout d'abord un état des lieux économique dans le domaine de l'agroalimentaire. Dans un second temps, il sera question de passer en revue les principaux moyens de compétitions qui sont à la disposition des entreprises performantes.
Par exemple, elles doivent tout d'abord gérer au mieux les aspects législatifs concernant les produits qu'elles souhaitent commercialiser. C'est un moyen de respecter les règles du jeu mais aussi de rendre un service supplémentaire au consommateur. C'est également, en partie les objectifs de la politique de qualité. Mais rassurer les consommateurs sur la qualité des produits agroalimentaires ne suffit pas à les rendre fidèles dans leurs achats. En effet, ces derniers sont très friands de changements, de nouveautés. Leurs besoins changent avec leur style de vie et inversement. Face à ces exigences, l'entreprise agroalimentaire doit s'adapter.
Innover devient donc de plus en plus nécessaire. Les consommateurs souhaitent et attendent des produits nouveaux et améliorés. L'innovation permanente semble être la seule façon d'éviter l'obsolescence de la gamme de produits d'une entreprise.

Or, il devient de plus en plus difficile de réussir le développement de nouveaux produits pour les raisons suivantes :

- manque d'idées nouvelles
- morcellement des marchés, des volumes de ventes et profits réduits
- durcissement de la législation et prise de conscience des consommateurs
- augmentation rapide des coûts de développement des nouveaux produits
- taux d'échec élevé lors de la commercialisation
- diminution de la durée de vie des produits commercialisés avec succès

Plusieurs solutions orchestrées soit par l'innovation, le marketing ou encore l'analyse sensorielle ont été proposées pour résoudre ce dilemme. En parallèle, l'étude du comportement du consommateur permet de déterminer l'ensemble des relations complexes et les facteurs d'influence qui caractérisent la consommation.
Chaque démarche est intéressante, complémentaire et apporte de réelles solutions de compétition. D'une manière globale, ces solutions existantes sont satisfaisantes pour l'entreprise. Mais aucune d'entre-elles ne s'intéresse vraiment à l'interaction directe entre le consommateur et le produit alimentaire au niveau de la perception sensorielle. C'est justement ce point faible que nous souhaitons aborder et développer dans le deuxième chapitre.
L'étude de la compétition en agroalimentaire à travers l'innovation et le comportement du consommateur est de nature complexe. Il a semblé important au Groupe DANONE, commanditaire de cette étude, de s'intéresser à la perception sensorielle des consommateurs vis à vis des produits alimentaires car elle est une voie possible d'innovation qui n'avait jamais encore été abordée.

Le deuxième chapitre a donc pour objet, de montrer comment les méthodes bibliométriques et lexicométriques nous ont aidés à travailler sur les commentaires libres de consommateurs. Aussi, nous détaillerons la méthode d'analyse des commentaires libres que nous avons mise en place après avoir, bien entendu, caractérisé les données, les objectifs de l'étude et analysé les méthodes existantes à partir desquelles nous avons commencé à travailler.

Le succès d'un produit dépend aussi beaucoup du plaisir qu'il procure aux consommateurs et donc en grande partie de ses qualités gustatives. Or, l'analyse sensorielle est spécialement destinée à examiner les propriétés organoleptiques d'un produit par les organes des sens.
C'est dans cet esprit que le projet de recherche sur la mise au point et la valorisation de nouvelles procédures pour mieux connaître la perception des produits alimentaires par les consommateurs s'est mis en place, car l'expression spontanée du consommateur est intéressante pour l'analyse sensorielle pour comprendre comment il ressent le produit et surtout comment il l'exprime.
Les commentaires libres de consommateurs peuvent être qualifiés de données textuelles au même titre que les entretiens, les discours ou tout texte littéraire. Mais ils ont tout de même leurs spécificités qui les rendent difficiles à traiter.
Si cette approche est nouvelle pour l'Analyse Sensorielle, elle est depuis longtemps abordée dans les disciplines littéraires. L'analyse des méthodes existantes en traitement des données textuelles peut nous aider à aborder nos travaux sous un nouvel angle.

De 1992 à 1996, trois méthodes de traitement des données textuelles (SPAD.T, ALCESTE et CANDIDE) ont été utilisées et ont chacune apportée des propriétés intéressantes pour le traitement des commentaires libres de consommateurs. Elles ont notamment confirmé la richesse de l'information textuelle mais des lacunes techniques nous ont incités à aller plus loin dans nos investigations sur les méthodes de traitement :

- un projet de collaboration a été établi avec un laboratoire spécialisé en bibliométrie, le Centre de Recherche Rétrospectives de Marseille afin de réaliser un échange de connaissances entre les méthodes de bibliométrie et de traitement des commentaires de consommateurs. Les premiers travaux ont consisté à éprouver les outils et méthodes du CRRM sur les données consommateurs.
- une évaluation de plusieurs logiciels du commerce a été effectuée.

En 1996 aucun outil n'a été choisi car bien qu'apportant tous certaines spécificités intéressantes, ils se révélaient être soit trop pointus, soit trop généraliste ou encore trop coûteux. Cependant, l'ensemble des méthodes et des outils étudiés dans cette partie apporte tous des résultats intéressants. Nous avons pu, grâce à ces différentes évaluations, préciser toutes les opportunités de traitements liés aux données textuelles pour déterminer l'approche idéale pour des données sensorielles. Trois axes principaux en découlent :

- amélioration du mode de collecte
- amélioration du codage
- simplification de la lecture des résultats

Les trois objectifs correspondent à des méthodologies connues et employées en sciences de l'information (notamment en bibliométrie) ainsi qu'en lexicométrie. En effet, l'application de la bibliométrie en veille technologique aborde fréquemment les techniques automatiques de codage et de représentation des données textuelles. De même en lexicométrie, la réduction de la dispersion du vocabulaire est très utilisée pour analyser les discours ou les œuvres littéraires. Nous nous en sommes donc inspirés pour mettre en place une nouvelle chaîne de traitement.

A travers la connaissance du consommateur, nous avions espéré au premier chapitre trouver des voies nouvelles dans la recherche d'innovation. Le deuxième chapitre nous a montré que le potentiel était là mais que la tâche ne serait pas simple pour le mener à bien.

Grâce à une confrontation d'idées en provenance de divers horizons et de compétences pluridisciplinaires, le troisième chapitre montre que nous sommes arrivés à mettre au point une méthode satisfaisante pour valoriser l'expression libre des consommateurs après dégustation de produit alimentaire.
La démarche adoptée nous semble tout à fait intéressante dans la mesure où nous sommes partis d'un problème très pratique sur la recherche de la connaissance du consommateur que nous avons essayé de mesurer grâce aux méthodes très théoriques des statistiques.
L'analyse des solutions existantes et proches de nos besoins nous a permis de mieux maîtriser les techniques et d'envisager d'autres solutions de traitement pour nos données. Nous avons trouvé un grand nombre d'idées nouvelles grâce à l'association de plusieurs domaines tels que l'analyse sensorielle, la lexicométrie, la veille technologique, la bibliométrie, …
Ce mélange de compétences a en définitive enrichi considérablement notre travail et a ouvert de nombreuses autres perspectives de recherche.
D'une manière globale, cette méthode s'inscrit dans une démarche classique de traitement de l'information telle qu'on peut l'envisager notamment en veille technologique. Nous retrouvons en effet les différentes étapes de collecte, traitement, d'analyse et validation, de diffusion et de capitalisation.
Ce parallèle permet également de montrer que notre démarche s'inscrit dans un processus d'intelligence économique. Or, la gestion stratégique de l'information est devenue l'un des moteurs essentiels de la performance globale des entreprises et des nations. En effet, le processus de mondialisation des marchés contraint les agents économiques à s'adapter aux nouveaux équilibres qui s'établissent entre concurrence et coopération. Désormais, la conduite des stratégies industrielles repose largement sur la capacité des entreprises à accéder aux informations stratégiques pour mieux anticiper les marchés à venir et les stratégies des concurrents.
Dans ce sens, grâce à la valorisation des commentaires libres de consommateurs et la maîtrise de ce type d'information, nous sommes arrivés à mieux connaître leur perception des produits alimentaires. Sa mise en pratique, développée dans le troisième chapitre, a pu démontrer son utilité dans une démarche globale d'innovation en agroalimentaire.

CHAPITRE I

Est-ce une évidence d'énoncer que pour conquérir des marchés, accroître ses bénéfices, se développer, toute entreprise doit ouvrir sa porte aux nouvelles technologies ?

Pour accéder facilement aux innovations, elle doit maîtriser les transferts de connaissances de la recherche fondamentale à la Recherche & Développement puis à la production. Elle doit également bien connaître les acteurs des transferts de technologies.

Aujourd'hui, l'apparition et le développement d'un produit ou d'un procédé font appel à des connaissances très variées sur le plan fondamental. Le passage de la connaissance fondamentale à la production n'est pas linéaire c'est-à-dire qu'un savoir n'aboutit pas à une innovation en passant par une phase de développement. C'est un complexe de savoirs issus de domaines différents qui peut conduire à un produit commercialisable. Ceci représente la partie amont de l'innovation.

Mais en aval, qu'est-ce qui va faire qu'un produit sera consommé ? Il y a là encore un enchevêtrement de circonstances qui vont déterminer la réussite de l'entreprise. Parmi elles, le consommateur y a une place de choix. Il s'agit d'un véritable indicateur de la bonne santé d'une entreprise. C'est pour cette raison que le marketing l'intègre dans ses études.
Dans le même ordre d'idée, l'analyse sensorielle s'intéresse à la préférence organoleptique des produits alimentaires par les consommateurs.

Parmi les différentes approches existantes pour aider les entreprises agroalimentaires à se développer, nous avons choisi de nous pencher sur la connaissance du consommateur.
Ce premier chapitre reprend donc l'état des lieux dans le domaine de l'agroalimentaire avec les principales stratégies de développement économique. Il permet de situer le contexte du projet de recherche réalisé au cours de cette thèse.

CHAPITRE I : CONTEXTE DE LA RECHERCHE

1. L'agroalimentaire

Le terme agroalimentaire est ambigu, car il désigne un domaine économique dont les contours sont variables. C'est pour cette raison que nous prendrons comme référence la définition de l'I.N.S.E.E.[1] : « Ensemble des entreprises transformant des produits en général d'origine agricole pour satisfaire les besoins alimentaires des consommateurs ».

Il y a trente ans en France, l'alimentation était placée dans les priorités du pays avec près de 26% des dépenses budgétaires (33,3% en 1960). Sa part dans les dépenses totales est tombée à 18% en 1995. Compte tenu du fait que la consommation de produits alimentaires a cru de 43% en volume au cours de ce quart de siècle, la consommation des français a tout de même augmenté de 1% en moyenne par an.

Dans le même temps, l'augmentation des prix a été moins forte que dans les autres secteurs de dépenses. Cette baisse relative des prix observée tout au long de la période s'explique en partie par des mouvements de la distribution et l'augmentation de la productivité sur certains produits agricoles **[MONC95]**.

Mais le phénomène le plus impressionnant durant ces 25 dernières années sur la fonction « alimentation » est sans doute la mutation de nos habitudes alimentaires ; en 1970, près des deux tiers de nos dépenses alimentaires étaient constitués de produits basiques peu ou pas transformés : produits agricoles, lait, viande et volaille, beurre, pain et sucre. L'augmentation de la consommation de ces produits basiques a été inférieure à celle de la croissance démographique. Aujourd'hui, ces produits représentent à peine 54% de notre consommation alimentaire totale. Par exemple, entre 1980 et 1995, la consommation de pommes de terre a baissé de 15,2% en volume et celle de la viande de 7,7% au profit des plats cuisinés. La consommation alimentaire de produits dits élaborés ou industriels a, quant à elle, doublé (90% de plus).

[1] Institut National de la Statistique et des Etudes Economiques

Nous avons donc assisté à l'émancipation des industries agroalimentaires vis à vis de l'agriculture, en associant nouveaux procédés de fabrication et nouveaux services aux consommateurs. Economiquement, leur taux de valeur ajoutée et leurs facteurs de développements se sont rapprochés de ceux des autres secteurs industriels (abaissement des coûts, politique de marque, innovations, …) **[INSE97]**.
Mais cette nouvelle orientation de notre consommation alimentaire s'explique aussi par :

- des facteurs technologiques et techniques tout d'abord avec la généralisation de l'équipement ménager : réfrigérateurs, cuisinières, fours électriques ou à gaz dans un premier temps : congélateurs dans un second temps et micro-ondes plus récemment.
- des facteurs sociologiques ensuite avec l'évolution du taux d'activité des femmes de 25 à 54 ans qui a doublé entre 1965 et 1990 pour dépasser aujourd'hui 80%. La progression de l'activité extérieure de la femme s'est traduite à la fois par la recherche de gain de temps (utilisation de produits industriels), une ouverture plus grande sur le monde extérieur, l'expérience de nouveaux goûts (repas pris à l'extérieur) mais également, en raison de l'urbanisation croissante, par la perte des traditions alimentaires et des compétences culinaires.
- d'autres facteurs économiques : le pouvoir d'achat accru, seulement pour partie avec le développement du travail féminin, a permis l'achat de produits plus élaborés et plus chers. Une tendance commune à tous les pays européens **[NEFU89]**.

Pendant que les industries agroalimentaires se modernisaient et augmentaient leur productivité, une mutation de la distribution se préparait. En effet, dans le même temps, le libre-service s'est généralisé et les grandes chaînes de distribution se sont constituées. Cette production de masse s'est accompagnée de changements qualitatifs par l'innovation produit : dessert lacté (yaourt), conserves, produits instantanés, produits surgelés, plats cuisinés, produits de quatrième gamme[2], etc…
Tout ceci a contribué à modifier considérablement les habitudes alimentaires des consommateurs puisqu'elle les a plongés dans ce que nous appelons « la société de consommation » **[MALA86]**.

Les paragraphes suivants vont nous permettre de détailler plus précisément la situation de l'agroalimentaire en France, en Europe et dans le monde sur le plan économique.

[2] Fruits et légumes prêts à l'emploi

1.1.1. Situation en France

Sur le marché national

Le poids de l'ensemble du secteur agroalimentaire dans l'économie nationale est actuellement d'environ 6% du produit intérieur brut marchand (PIBM). Il représente, en emplois et en valeur ajoutée, environ 1/17ème (soit 6%) de l'économie française **[COHE96]** et **[COHE97]**.

Le secteur des industries agroalimentaires est le premier secteur industriel français et le premier secteur excédentaire de la balance commerciale : 45,8 milliards de francs en 1995 **[MONC97]**. Ces industries transforment environ 50**%** de la production agricole française et contribuent plus largement à la formation du PIBM que l'agriculture (198 milliards de francs [3,6%] contre 156,9 milliards de francs [2,4%], en 1994). Près de 4 200 entreprises de plus de 10 salariés contribuent à l'activité de ce secteur et dégagent un chiffre d'affaires de 647 milliards de francs par l'emploi de 397 000 personnes. Les principaux produits concernés sont les viandes (24,3%), les produits laitiers (20,8%), l'épicerie (20,6%, biscuiteries, confiseries, boulangerie, sucre, etc. ...) et les boissons (14,1%).

Quel que soit le secteur, les industries agroalimentaires possèdent une très grande variété de types d'entreprises. D'autre part, 95% des entreprises de plus de dix salariés sont représentées par des petites et moyennes entreprises[3]. Pourtant, elles réalisent 63% du chiffre d'affaire sur le territoire français et leur poids en valeur ajoutée n'est que de 52%. Le Tableau 1 montre que les grands groupes s'octroient la plus grande part du marché et évidemment les meilleures places.

[3] Moins de 500 salariés

Tableau 1 : Les 20 meilleurs chiffres d'affaires des industries alimentaires en France en 1996

Sociétés	Activités	CA* (1995)	CA* (1996)
DANONE (Groupe)	Multi spécialités	79.5	83.9
ERIDANIA BEGHIN-SAY	Multi spécialités	50.8	55
NESTLE (Nestlé France + Perrier Vittel)**	Multi spécialités	40.8	42
BESNIER (Cie Laitière)	Industrie laitière	25	26.1
SOUFFLET (Ets Jean)	Meunerie	16.9	17
PERNOD RICARD	Multi spécialités dans les boissons	15.9	16.8
SODIAAL	Industrie laitière	16.5	16.5
SOCOPA SA	Abattage de bétail	11.3	12
COMPAGNIE LAITIERE EUROPEENNE (CLE)	Industrie laitière	11.5	11.4
MOET HENNESSY** (LVMH SPIRITUEUX)	Multi spécialités dans les boissons	11.1	11.3
UNILEVER FOOD**	Multi spécialités	9.9	11
BONGRAIN SA	Industrie laitière	9.9	10.4
COOPAGRI BRETAGNE (CAB)	Multi spécialités	8.3	8.8
BEL	Industrie laitière	8	8.6
MARS (Groupe)**	Multi spécialités	8.3	8.4
CASTEL FRERES (Groupe)**	Multi spécialités dans les boissons	8	8.3
CANA (Groupe)	Multi spécialités	8.3	8.2
KRAFT JACOBS SUCHARD France	Chocolat, confiseries	8.2	7.8
DOUX	Abattage et transformation de volailles	7.5	7.7
REMY COINTREAU	Multi spécialités dans les boissons	6.9	7

(Source AGRA ALIMENTATION - novembre 1997)
* En milliard de francs
** Estimation

Sur le marché international

Plusieurs faits interviennent dans la difficulté à entrer dans la compétition pour l'agroalimentaire français :

- la standardisation de la consommation alimentaire : l'unification européenne a provoqué une homogénéisation des produits élaborés.
- l'influence de la grande distribution : les entreprises agroalimentaires ont dû s'adapter aux exigences des distributeurs (respect des délais de livraison, homogénéité des lots,…). De plus, ces derniers intègrent de plus en plus des produits très compétitifs sous leur marque (voir Tableau 2). Ceci a pour effet de freiner la progression du prix.
- la concurrence internationale : peu de grands groupes agroalimentaires français sont situés dans les premières places (voir Tableau 3 en page 8). Le marché est en effet largement dominé par les multinationales américaines **[EURO91]**. De la même façon, sur le plan européen, la France se positionne dernière les Iles Britanniques (voir Tableau 4 en page 12)
- le manque d'innovation : les dépenses de recherche et d'innovation sont très en dessous de ceux de leurs grands compétiteurs (voir Tableau 5 en page 20).

Tableau 2 : Poids des marques de distributeurs (MDD) en France en 1996

Enseignes	MDD (en % des valeurs)
Casino	24,8
Intermarché	24,7
Géant	20
Carrefour	18,9
Système U	18,5
Continent	17,8
Stoc	16,2
Auchan	15,7
Champion	15,1
Centres Leclerc	14,8
Cora	12,2
Atac	11
Mammouth	8,9

Pourtant, l'agroalimentaire est un secteur clé pour notre pays sur le plan international car il occupe toujours une position de leader. En effet, les derniers résultats du commerce extérieur affiche un excédent de 4,8 milliards de francs à la fin du premier semestre 1997. Il est en très forte augmentation par rapport à 1996. Cette progression des exportations (8 %) témoigne du dynamisme des industries alimentaires et de l'internationalisation croissante de cette activité. Corollaire de cette progression, l'industrie alimentaire est très sensible aux conditions qui régissent l'Organisation Mondiale du Commerce. Elle est de ce fait soumise à de fortes contraintes tant du point de vue de la concurrence intérieure et extérieure, que des pressions de la grande distribution **[VALM95]** et de l'opinion publique (cf. Affaire de la vache folle).

1.1.2. Situation dans le monde

Dans les pays en voie de développement

Aujourd'hui, environ la moitié des habitants de la planète vivent dans une zone de précarité alimentaire.
Or, l'augmentation des ressources alimentaires mondiales dépend fortement de la disponibilité en diverses ressources de base à savoir les terres non encore mises en valeur, l'eau pour l'irrigation, l'énergie, les fertilisants, les pesticides et autres ressources indispensables pour intensifier la production agricole.
Nous nous trouvons donc face à un grave et difficile problème qui est l'impossibilité de subvenir aux besoins alimentaires du globe.
Récemment, des solutions nouvelles laissent entrevoir un certain espoir. En effet, les actions de contrôle de la démographie combinées aux applications du génie génétique pourraient associer la diminution de la population dans certaines régions et l'augmentation des ressources alimentaires indépendamment des facteurs limitant vus ci-dessus **[FAOG98]**.

Dans les pays développés

Le commerce international est de plus en plus structuré par l'existence de groupes multinationaux (voir Tableau 3 et Tableau 4) **[LEBR96]**. La consommation des pays développés a tendance à diminuer à cause du vieillissement de la population (baisse de la natalité, augmentation de l'espérance de vie…), de la réduction des revenus (chômage, allongement de la durée des études,…) et de la disparition de la cellule familiale (divorces) **[CASE97]**.

La conquête des marchés émergeants représentent sans conteste un moyen d'augmenter leur croissance. Mais les éléments vitaux garantissant le succès au sein des marchés existants restent tout de même l'innovation de produits et de procédés ainsi qu'une meilleure connaissance du consommateur.

Tableau 3 : Les 20 premiers groupes alimentaires dans le monde en 1995

Sociétés	Pays	Ventes*(1995)
Nestlé	Suisse	38,8
Philip Morris Companies Inc.	Etats-Unis	33,38
Unilever Plc/NV	Royaume-Uni/Pays-Bas	26,76
ConAgra Inc.	Etats-Unis	24,82
PepsiCo Inc.	Etats-Unis	19,09
Coca-Cola Company	Etats-Unis	18,02
Danone	France	14,19
Archer Daniels Midland Co.	Etats-Unis	13,31
Mars Inc.	Etats-Unis	13,00
Grand Metropolitan Plc	Royaume-Uni	12,69
IBP Inc.	Etats-Unis	12,54
Kirin Brewery Co. Ltd.	Japon	11,56
CPC International Inc.	Etats-Unis	9,84
Anheuser Busch Co. Inc.	Etats-Unis	9,59
Sara Lee Corp.	Etats-Unis	9,43
Associated British Foods Plc	Royaume-Uni	9,21
H.J. Heinz Company	Etats-Unis	9,11
Asahi Breweries Ltd.	Japon	9,11
Eridania Béghin Say	France	9,07
R.J.R. Nabisco, Inc.	Etats-Unis	8,29

**de produits agroalimentaires en milliards de DUS*

Tableau 4 : 10 premiers groupes alimentaires en Europe en 1996 rangés par ordre des ventes mondiales en millions de USD

Sociétés	Pays	Ventes alimentaires	Ventes totales	Bénéfices
Nestlé SA	Suisse	48 231	50 241	5 192
Unilever Plc/NV	Royaume-Uni/Pays-Bas	25 078	51 179	2 458
Danone	France	13 911	15 121	616
Grand Metropolitan	Royaume-Uni	12 057	13 701	592
Eridania Beghin-Say	France	9 999	9 999	300
Cadbury Schweppes	Royaume-Uni	7 809	7 809	904
Heineken NV	Pays-Bas	7 338	7 338	64
Guinness Plc	Royaume-Uni	7 221	7 221	974
Dalgety Plc	Royaume-Uni	6 619	6 619	90
Tate & Lyle Plc	Royaume-Uni	6 460	7 878	315

Source: Datamonitor Global Food & Drink Companies Database

2. Moyens de compétition sur le produit

Les industries agroalimentaires se trouvent donc à l'heure actuelle devant un marché saturé. Elles doivent s'armer de nouvelles solutions de compétition pour rester dans la course ou pour se positionner parmi les premiers.
Elles ont d'abord pensé à se réorganiser. Pendant que leur principale activité d'acquisition était concentrée sur les marchés émergeants (Asie, Amérique latine, Europe de l'Est …), les grands groupes agroalimentaires opéraient des actions de recentrage et de regroupement au sein de leurs activités. Ces opérations leur ont permis de se renforcer sur leurs spécialités pour faire face en partie, à la nouvelle concurrence de la grande distribution.
Elles doivent aussi respecter les réglementations en vigueur, suivre les nouvelles tendances et promouvoir la qualité de leur produits pour respecter leurs clients et les inciter à leur faire confiance.
Elles peuvent innover. L'innovation de produits représente la moitié des marques d'une entreprise compétitive. Ce choix est également très risqué sur le marché actuel. C'est pourquoi il est réservé la plupart du temps aux grands groupes. Certains se contentent même d'imiter des produits innovants pour limiter les échecs.

Dans tous les cas, elles ont tout intérêt à mieux connaître le consommateur. Le consommateur est bien sûr la clé du succès pour introduire un nouveau produit. En effet, il est impensable de vendre un nouveau produit que personne n'achètera. Mais comment connaître les attentes des grandes masses de consommateurs dont les goûts diffèrent en fonction de leurs pays, de leurs cultures, et de bien d'autres facteurs ? Il ne s'agit plus de proposer des produits génériques destinés à des segments de marchés hypothétiques. Aujourd'hui, le consommateur désire des produits nouveaux, différents et agréables dans la mesure où ils sont adaptés à son style de vie et à son goût.
D'une certaine manière, l'enjeu de l'entreprise agroalimentaire est lié à son degré de flexibilité et d'agilité pour produire de façon interactive et en fonction de la demande des consommateurs. Ceci est aujourd'hui envisageable grâce aux nouvelles technologies de l'information et de la communication (datamining, réseaux informatiques, groupware, …).

Parmi les stratégies que nous venons d'énoncer, il nous semble important de détailler les aspects suivants :

- la réglementation
- la qualité
- l'innovation

Ces trois points sont importants pour comprendre la problématique générale du développement des nouveaux produits alimentaires. Dans ce contexte, il sera plus aisé d'expliquer le cheminement de notre démarche vers la recherche de la connaissance du consommateur.

2.1. Gestion des réglementations

2.1.1. Importance des normes

Nous venons de voir que les industries agroalimentaires françaises occupent une place de premier plan dans l'économie nationale (voir section 1.1.1). Ce secteur a donc une importance stratégique autant sur le plan national qu'international. Cependant, il est sans cesse menacé par la guerre économique.

Dans ce contexte, le recours à la normalisation par le secteur agroalimentaire représente un enjeu majeur **[MILL94]**.
Depuis sept années, le Comité d'Orientation Stratégique (COS) agroalimentaire de l'AFNOR travaille à l'élargissement de la normalisation à de nouvelles perspectives **[EINA97]**.
Hier consacrée essentiellement aux méthodes d'analyse, la normalisation en agroalimentaire concerne aujourd'hui dans près de 30 % des cas, des dénominations de produits, spécifications, guides de bonnes pratiques de production et de transformation et la description de méthodes de travail qui sont des instruments d'organisation des entreprises et des laboratoires (par exemple la traçabilité dans le domaine des viandes, le protocole d'élaboration d'un guide hygiénique ...).
Nous pouvons donc considérer aujourd'hui que l'acclimatation de la normalisation aux industries agroalimentaires s'est renforcée et que les normes représentent clairement un élément de référence souvent déterminant dans les relations commerciales.

Elle se traduit d'abord par l'utilisation des normes dans les stratégies professionnelles. La normalisation est en effet déterminante pour lutter contre les risques de concurrence déloyale sur les produits, pour assainir le marché en favorisant l'émergence d'une référence collective reconnue - la norme - dans les relations contractuelles. Elle offre en outre la possibilité de segmenter le marché ainsi que certains secteurs l'ont démontré par leur usage des normes.
La normalisation sert aussi de base de référence pour les démarches de certification de manière directe quand il s'agit de la marque NF Agro ou de manière indirecte comme outil de mesure.

L'harmonisation des méthodes d'analyse de référence représente un enjeu majeur pour l'ensemble des acteurs économiques. En effet, la diversité de méthodes peut engendrer une entrave aux échanges et conduire à une duplication des contrôles se traduisant par une augmentation globale de leur coût.

La sécurité des aliments, exigence impérative par excellence, fait toujours l'objet de nombreux débats (cas de l'Encéphalopathie Spongiforme Bovine[4]).
L'enjeu consiste à fournir des documents de référence reconnus, et harmonisés sur lesquels pourront s'appuyer les entreprises. En effet, la directive hygiène 93/43 confirme que leur responsabilité est engagée et ce texte les incite à développer des instruments volontaires - guides de bonnes pratiques hygiéniques - éléments de preuve de leur maîtrise de l'hygiène.
Après avoir défini le cadre méthodologique de l'élaboration de ces guides, la Commission de Normalisation a mis en œuvre l'établissement d'un document de référence terminologique afin de lutter contre des interprétations qui, de par leurs divergences, sont des sources d'entrave et de distorsion de concurrence.

En parallèle, poursuivant son approche méthodologique, la commission de normalisation travaille à l'élaboration d'un guide décrivant comment établir un protocole d'évaluation de la date limite de consommation des différents produits alimentaires.

2.1.2. Autres réglementations

La Direction générale de la concurrence, de la consommation et de la répression des fraudes au ministère des Finances a entre autres la mission de contrôler la qualité (et la sécurité) des denrées agricoles ou alimentaires mises sur le marché **[MULT91]**.
D'une manière générale, la présentation d'un produit alimentaire est assujettie à un certain nombre d'obligations positives (nécessité de renseigner l'acheteur sur les caractéristiques essentielles du produit, sur la quantité, éventuellement sur ses effets utiles) et d'obligations négatives (interdiction des mentions fausses ou induisant en erreur). Ces obligations sont précisées par des textes horizontaux, comme le décret du 7 décembre 1984 concernant l'étiquetage et la présentation des denrées alimentaires ou des règles verticales, propres à chaque denrée ou catégorie de denrées, les uns comme les autres pris en application de la loi du 1er août 1905.

[4] Maladie de la vache folle

2.1.3. Evaluation scientifique et demande d'autorisation de mise sur le marché

Depuis le 15 mai 1997, les "nouveaux aliments" et les "nouveaux ingrédients alimentaires" (novel foods) font désormais l'objet d'une procédure d'évaluation scientifique préalable à leur mise sur le marché et d'un étiquetage obligatoire **[MINI98]**. Sont notamment concernés :

- les aliments et ingrédients alimentaires contenant des organismes génétiquement modifiés ou consistant en de tels organismes,
- les aliments et ingrédients alimentaires produits à partir d'organismes génétiquement modifiés, mais n'en contenant pas,
- les aliments et ingrédients alimentaires présentant une structure moléculaire primaire nouvelle ou délibérément modifiée,
- les aliments et ingrédients alimentaires composés de micro-organismes, de champignons ou d'algues ou isolés à partir de ceux-ci,
- les aliments et ingrédients alimentaires isolés à partir d'animaux, à l'exception des aliments et ingrédients alimentaires obtenus par des pratiques de multiplication ou de reproduction traditionnelles et dont les antécédents sont sûrs en ce qui concerne l'utilisation en tant que denrées alimentaires,
- les aliments et ingrédients alimentaires auxquels a été appliqué un procédé de production qui n'est pas couramment utilisé, lorsque ce procédé entraîne dans la composition ou dans la structure des aliments ou ingrédients alimentaires des modifications significatives de leur valeur nutritive, de leur métabolisme ou de leur teneur en substances indésirables.

Pour mettre sur le marché de tels produits, ils doivent être autorisés. Tout demandeur doit donc désormais fournir un dossier technique et scientifique à l'Etat membre[5] dans lequel le produit sera mis sur le marché pour la première fois et transmettre une copie simplifiée de son dossier à la Commission européenne. Une procédure d'expertise, qui est plus ou moins complexe et plus ou moins centralisée par Bruxelles, selon la nature du produit, se met en route. La première finalité de cette procédure : s'assurer, notamment, que le produit est sans danger pour le consommateur et qu'il n'implique pas d'inconvénients nutritionnels. La seconde, permettre à tous les autres Etats membres, qui le souhaitent, d'exprimer leurs objections ou leurs exigences complémentaires.

Pour être accordée, les autorités européennes ont prévu que l'autorisation de mise sur le marché devait indiquer clairement la dénomination de l'aliment ou de l'ingrédient alimentaire visé, ses spécifications et ses conditions d'utilisation. En outre, l'aliment ou l'ingrédient alimentaire ne peut être commercialisé que si son étiquetage comporte obligatoirement, en plus des mentions obligatoires habituelles, des précisions pour informer le consommateur.

Les modalités précises de cet étiquetage sur la forme et le contenu font actuellement l'objet, en France, de réflexions au sein du Conseil National de l'Alimentation (CNA). Le groupe de travail en charge du dossier préconise de *"traiter séparément"* les organismes génétiquement modifiés et les produits issus d'organismes génétiquement modifiés (c'est à dire ne contenant plus d'acide désoxyribonucléique (ADN) recombinant biologiquement actif). Ainsi, à son sens, l'étiquetage des tomates génétiquement

[5] de la Communauté Européenne

modifiées mises en vente devrait être obligatoirement complété de la mention *"génétiquement modifiées"*. Par contre, l'étiquetage des huiles, issues de tournesols modifiés pour être plus riches en acides gras insaturés, devrait être obligatoirement complété de la mention *"issues d'organismes génétiquement modifiés"*.

Les industries agroalimentaires n'ont finalement pas le choix. Elles doivent veiller à gérer au mieux les aspects législatifs concernant les produits qu'elles souhaitent commercialiser. C'est un moyen de respecter les règles du jeu mais aussi de rendre un service supplémentaire au consommateur. C'est également en partie les objectifs de la politique de qualité.

2.2. Promouvoir la qualité des produits

La politique de qualité regroupe à la fois l'identification et la garantie de la qualité des produits. Elle constitue un enjeu considérable pour le secteur agricole et alimentaire **[SYLV92]**. Plusieurs constats justifient cette analyse.

Tout d'abord, la construction des règles du commerce international (dans le cadre de l'organisation mondiale du commerce ou du marché unique européen), se fonde tout particulièrement sur la lutte contre les entraves techniques aux échanges afin de permettre la libre circulation des denrées alimentaires.
Ce choix provoque des changements fondamentaux dans l'élaboration du droit alimentaire, le rôle des services officiels de contrôle et la responsabilité des entreprises quant au respect de la réglementation et à la qualité des produits mis en vente.
La réglementation s'attache aujourd'hui à ne fixer que ce qui est strictement nécessaire au fonctionnement du marché tout en garantissant un haut niveau de protection du consommateur.
Elle se limite aux exigences impératives que sont la protection de la santé et la sécurité du consommateur, son information, la loyauté de la concurrence, la protection de l'environnement et l'harmonisation des contrôles. Elle fixe des exigences de résultats laissant aux entreprises le choix des moyens, notamment à travers l'utilisation de guides de bonnes pratiques.
Parallèlement, la Cour de Justice Européenne a fixé des limites précises à la restriction de circulation de toute denrée alimentaire légalement produite dans un Etat membre de la Communauté Européenne. Le principe de base est la reconnaissance mutuelle des règles non harmonisées au niveau communautaire : un produit loyalement fabriqué et commercialisé dans un Etat membre de la Communauté peut être commercialisé dans tous les autres Etats membres.

Cette évolution se traduit par l'introduction d'une plus grande souplesse favorable à l'innovation et par une plus forte responsabilisation des opérateurs quant à la qualité des produits mis en marché. Mais baser l'information du consommateur sur le seul étiquetage peut engendrer des distorsions de concurrence et une tromperie du consommateur, en particulier pour les produits basiques de première transformation ou pour les produits traditionnels pour lesquels :

- le temps passé par le consommateur pour l'acte d'achat est très court,
- la perception du niveau qualitatif du produit n'est pas possible par la seule lecture de la liste des ingrédients.

Dans ces conditions, nous pouvons craindre une grande anarchie sur le marché avec dégradation de la qualité et concurrence déloyale entraînant progressivement une désaffection du consommateur pour ces produits.
C'est pour cela que se sont mis en place des outils volontaires, d'identification, de certification et de protection de qualité : la certification de système d'assurance de la qualité, l'accréditation des laboratoires, le recours à la normalisation tant pour la reconnaissance de méthodes d'analyse de référence que pour des spécifications de produits, à la certification de produit. Ce sont autant d'outils mis à disposition des opérateurs pour organiser le marché et garantir la qualité de leurs produits et de leurs prestations.

D’autre part, nous avons vu (voir section 1) que le marché européen des produits agricoles et alimentaires est aujourd'hui globalement saturé. La concurrence est vive, voire parfois déloyale. Il est donc nécessaire de sortir d'une logique d'offre et d’entrer dans une logique de réponse à la demande, c'est-à-dire dans une démarche de qualité.
Il ne s'agit pas de fournir un produit standard où la différence se fait par le prix mais de segmenter le marché par des produits dont la qualité est identifiée, garantie et répond aux attentes du consommateur.
Les signes officiels de la qualité sont des outils mis à dispositions des opérateurs économiques pour segmenter le marché et assurer une concurrence loyale. Ils ont des moyens pour maintenir, voire créer de la valeur ajoutée.

Enfin, la consommation des produits agroalimentaires était auparavant très dépendante des catégories socioprofessionnelles. Aujourd'hui, une même personne peut acheter le même jour au même endroit des produits premiers prix et des produits haut de gamme. Elle cherche à se rassurer sur le mode d'obtention de ces produits et leur origine, étant donné la complexification des filières de production, et souhaite des produits plus authentiques et ayant plus de goût.
Il est donc indispensable de donner au consommateur les moyens d'identifier et de distinguer les produits qui bénéficient de qualités particulières : goût, origine géographique, savoir-faire, mode de production, et lui permettre de choisir en toute connaissance de cause.
C'est pourquoi les signes officiels de la qualité ont été mis en place : ils apportent la garantie officielle des Pouvoirs Publics sur la qualité et l'origine géographique des produits.
Cela ne signifie pas que les produits qui ne bénéficient pas de reconnaissance officielle de la qualité ne sont pas des produits de qualité mais ces signes apportent la confiance indispensable au consommateur.

Il existe quatre signes distinctifs : l'appellation d'origine contrôlée, le label rouge, la certification de conformité et l'agriculture biologique, qui ont chacun leur vocation particulière :

- l'A.O.C. permet la reconnaissance d'un produit typé souvent de grande notoriété et qui tire ses qualités de son terroir,
- le label rouge est la garantie d'un produit de qualité supérieure à celle des produits courants,
- la certification de conformité garantit que le producteur s'engage sur des caractéristiques et des règles de fabrication et assure la constance de la qualité de son produit,
- l'agriculture biologique recourt à des pratiques culturelles et d'élevage soucieuses de l'environnement et du bien-être des animaux.

Ces quatre signes trouvent leur prolongement direct dans la réglementation européenne qui permet d'assurer la protection juridique des dénominations de produits liées à une origine géographique ou issues d'un mode de production traditionnel (AOP, IGP, attestation de spécificité) ou du mode de production biologique.

Rassurer les consommateurs sur la qualité des produits agroalimentaires ne suffit pas à les rendre fidèles dans leurs achats. En effet, ces derniers sont très friands de changement, de nouveauté. Leurs besoins changent avec leur style de vie et inversement. Face à ces exigences, l'entreprise agroalimentaire doit s'adapter.

2.3. Innover

Les études prospectives pour les années à venir prévoient une baisse de consommation pour l'alimentation tandis que la progression du marché en volume devrait augmenter à une vitesse légèrement supérieure à la croissance démographique. Comment rétablir l'équilibre ?
Une réponse possible pour les industriels réside dans leur capacité à créer des produits à plus forte valeur ajoutée, qui apporteront au consommateur la meilleure qualité de service (traçabilité, qualité nutritionnelle, etc...) **[KERI93]**.

2.3.1. La veille

Avant de créer des nouveaux produits ou procédés, l'entreprise agroalimentaire a besoin de connaître et comprendre son environnement concurrentiel. Pour cela, elle doit collecter toutes les informations pertinentes, de nature formelle ou informelle, issues du monde scientifique, technique, technologique, économique, juridique, …
Pour connaître le milieu dans lequel elle évolue, l'entreprise doit donc rechercher cette information, l'analyser pour la comprendre, la synthétiser et la diffuser à ses acteurs décisionnels afin de les aider à prendre les meilleures décisions **[DOUH95]**. Pour réussir une telle organisation, il lui est nécessaire de développer un système d'intelligence économique avec des processus tels que la veille scientifique, technologique, économique, stratégique, …

Le processus de veille permet à la fois de surveiller l'environnement concurrentiel de l'entreprise, d'alimenter la stratégie mais aussi et surtout de ne pas refaire ce qui existe déjà.
L'un des éléments qui distingue l'innovation de la copie est cette capacité du chercheur d'obtenir, d'analyser et d'utiliser rapidement l'information.

Enfin, la veille produit consiste à surveiller les marchés et à dénicher les innovations aux quatre coins de la planète. Elle doit être le moteur d'un processus permanent d'écoute et de compréhension des enjeux majeurs pour l'entreprise, permettant une meilleure approche du client, en débanalisant l'offre et en anticipant les besoins **[MART89]**.

2.3.2. La Recherche & Développement

La recherche publique demeure importante même si les grandes entreprises, conscientes de la nécessité de développer l'investissement intellectuel et le soutien à l'innovation pour préparer l'avenir mettent, elles aussi, en place des structures de recherche **[MINI97]**.
En 1995, les crédits publics de recherche intéressant le domaine de l'agriculture et de l'agroalimentaire étaient de 3,6 milliards de francs. Parmi les principaux bénéficiaires, figurent l'institut national de la recherche agronomique **[INRA98]**, le centre national du machinisme agricole, du génie rural, des eaux et des forêts **[CEMA98]**, le centre national d'études vétérinaires et alimentaires **[CNEV98]** et l'institut français de recherche et d'exploitation de la mer **[IFRE98]**, les établissements d'enseignement supérieur et de recherche et les centres techniques fédérés au sein de l'association de coordination technique des industries agricoles (ACTIA) et l'association de coordination technique de l'agriculture (ACTA). En outre, ces crédits publics soutiennent certains programmes incitatifs (par exemple "Aliment demain") visant à favoriser la coopération entre industriels, professionnels, laboratoires publics et centres techniques.
Face aux évolutions que traversent l'agriculture et la société rurale, dans un contexte mondial plus concurrentiel, le Ministère de l'agriculture et de la pêche et le Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche s'emploient à développer une politique de recherche cohérente : en dotant les structures de recherche de moyens financiers, en modernisant le dispositif alliant la recherche fondamentale et la recherche appliquée, en associant les partenaires économiques et professionnels. Nous pouvons citer, à cet égard, les centres de recherche en nutrition humaine [CRNH], organisés autour d'un thème et regroupant des spécialistes en matière de santé de l'homme, de connaissance de l'aliment, de métabolisme cellulaire, … ou encore le Centre Européen des Sciences du Goût où se rejoignent le CNRS et l'industrie agroalimentaire sur des aspects neurologiques, biologiques, physiologiques, psychologiques, culturels et sociaux **[CNRS98]**.

En France, les industries agroalimentaires consacrent entre 0.25% et 0.3% de leur chiffre d'affaires et 1% de la valeur ajoutée à leur Recherche et Développement. A titre d'exemple, le groupe DANONE dépense 0.9% de son chiffre d'affaires en Recherche et Développement avec un effectif de 1 200 personnes **[DANO98]**.
Peu d'industries agroalimentaires font de la recherche et du développement et le budget qui lui est consacré est faible en comparaison d'autres secteurs industriels.
Il existe même des différences entre les secteurs de l'agroalimentaire. L'industrie de la viande fait très peu de recherche par rapport à l'industrie du lait.

L'importance de la recherche varie également en fonction de la taille de l'entreprise puisque les PME/PMI n'ont pas de structure spécifique à la recherche contrairement aux grands groupes.
La faiblesse de la recherche privée dans les industries agroalimentaires peut être expliquée par :

- la taille des industries. La majorité d'entre elles sont des PME-PMI avec peu de cadres techniques qui pourraient dialoguer avec la recherche publique. Il faut souligner cependant le rôle des centres techniques et professionnels auxquels les industriels peuvent faire appel pour résoudre des problèmes techniques.
- la recherche publique joue aussi un rôle important
- le risque commercial d'une innovation réelle, le consommateur étant généralement attiré par les produits qu'il connaît,
- le transfert de technologies se réalise via d'autres secteurs industriels (équipementiers, chimistes, …)

En France, même si l'industrie agroalimentaire est le premier secteur industriel, les entreprises consacrent une faible part de leur chiffre d'affaires à la recherche, au développement et à l'innovation. Ceci est en partie compensé par les moyens dont dispose la recherche publique **[MINI97]**.

Tableau 5 : Les chiffres de la recherche en France

Recherche et développement (RD)				
	1970	**1980**	**1990**	**1996**
Nombre de demandes de brevets d'invention déposées par la France	14 106	11 000	12 378	12 916
Nombre de marques françaises déposées	18 331	36 581	67 771	61 808

	1973	**1980**	**1990**	**1994**
Effectifs des chercheurs et ingénieurs en R&D	62 700	74 900	124 000	149 200
dont : en entreprises (%)	44,7	44,7	46	44,7

Dépense intérieure de recherche et développement (DIRD) :				
	1980	**1985**	**1990**	**1995**
en milliards de Francs	51	105,9	157,2	179,4
en % du PIB	1,8	2,25	2,4	2,3

Source : La France en bref, INSEE

2.3.3. Le brevet

Souvent dans le langage courant, les termes d'innovation et invention sont confondus. L'invention est plutôt assimilée à une idée nouvelle alors que l'innovation est une mise en pratique de cette idée, c'est l'aboutissement commercial ou industriel d'une invention. L'invention est la plupart du temps issue d'un travail de recherche. Pour la valoriser et la protéger, la propriété industrielle offre un ensemble de moyens juridiques, techniques et administratifs et le brevet en est une application.

En effet, le brevet peut permettre à l'entreprise à la fois d'éviter de travailler sur des idées qui ont déjà été développées et publiées de façon à ne pas être accusée de contrefaçon, de faciliter la résolution de certains problèmes techniques, de s'informer sur les produits en préparation de la concurrence et d'aider à obtenir la protection la plus sûre d'une invention **[JAKO94]**.

Mais ce n'est ni un bon indicateur d'innovation ni un bon indicateur d'invention ou de l'activité de la recherche. Il permet d'estimer le potentiel d'invention et d'innovation d'une entreprise mais il ne mesure que partiellement l'activité innovante et de recherche **[KABL94]**. Par contre, la marque indique directement la présence d'un nouveau produit sur le marché (voir section 2.3.5).

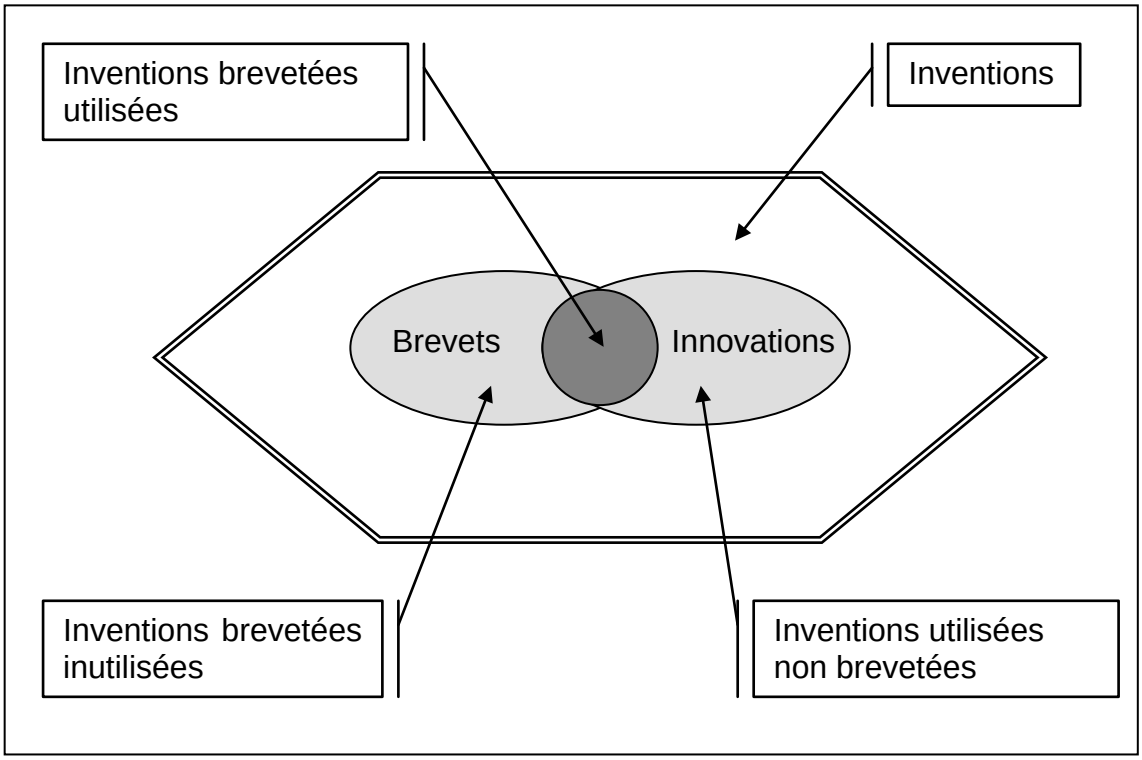


Figure 1 : Schéma des relations Inventions, Innovations, Brevets [BASB87]

La Figure 1 montre que toutes les innovations et les inventions ne sont pas brevetées. Il est facilement compréhensible que l'entreprise adopte une stratégie différente suivant l'enjeu qu'une invention pourra engendrer **[GUEL94]**. Certaines entreprises ne sont pas toujours à même d'évaluer la portée de leur invention. Il leur est alors difficile de déposer un brevet à ce stade. D'autres auront tendance à breveter des inventions sans avoir l'intention de les exploiter mais dans le seul but de freiner leurs concurrents.

Enfin, il existe également des entreprises qui préfèrent taire leurs découvertes pour surprendre leurs concurrents ou conserver leur savoir-faire difficilement décelable dans les produits finis.

2.3.4. Le produit agroalimentaire

Tableau 6 : Nombre de nouveaux produits agroalimentaires par catégorie aux Etats-Unis de 1989 à 1996

Catégories alimentaires	1989	1990	1991	1992	1993	1994	1995	1996
Aliments pour bébé	53	31	95	53	7	45	61	25
Pâtisserie	1 155	1 239	1 631	1 508	1 420	1 636	1 855	1 340
Ingrédients pour pâtisserie	233	307	335	346	383	544	577	419
Boissons	913	1 143	1 367	1 538	1 842	2 250	2 854	2 003
Céréales de petit déjeuné	118	123	108	122	99	110	128	121
Sucrerie/Chewinggum/Snacks	1 355	1 486	1 885	2 068	2 043	2 450	2 462	2 310
Condiments	1 701	2 028	2 787	2 555	3 147	3 271	3 698	2 815
Produits laitiers	1 348	1 327	1 111	1 320	1 099	1 323	1 614	1 345
Desserts	69	49	124	93	158	215	125	100
Entrées	694	753	808	698	631	694	748	597
Fruits & Légumes	214	325	356	276	407	487	545	552
Aliments pour animal	126	130	202	179	276	161	174	121
Viandes	509	663	798	785	453	565	790	637
4° gamme	489	538	530	560	680	980	940	611
Soupes	215	159	265	211	248	264	292	270
Totaux	9 192	10 301	12 398	12 312	12 893	15 006	16 863	13 266

Source: New Product News/PF.

Chaque année, de nouveaux produits agroalimentaires apparaissent sur le marché (voir Tableau 6) mais très peu constituent une réelle innovation résultant d'efforts de recherche importants. Il s'agit plutôt d'adaptation de recettes ou des changements de présentation. Nous avons affaire à de l'innovation incrémentale et non à de l'innovation radicale, c'est-à-dire à des améliorations de procédés ou de recettes déjà existantes plutôt qu'à des changements radicaux dans les procédés de fabrication ou qu'à l'apparition de produits inexistants sur le marché jusqu'à présent.

Tableau 7 : Les différentes sources d'innovations pour les entreprises

Facteur d'innovations comptant moyennement ou beaucoup pour l'entreprise	Pourcentage d'entreprises
Utilisation novatrice de biens d'investissements	54
Etudes de l'entreprise	46
Recherche et développement	37
Matériaux nouveaux	30
Recherche et développement extérieurs	26
Recherche et développement du groupe	25
Brevets extérieurs	8
Brevets de l'entreprise	7

Le Tableau 7 montre que l'utilisation de biens d'investissements est une source importante d'innovations pour les industries agroalimentaires alors que les brevets comptent peu. Dans d'autres secteurs comme la chimie, les brevets sont une source importante d'innovations.
Les entreprises agroalimentaires optent dans la plupart des cas pour l'exploitation (achat de licences, coopérations, …) de nouveautés technologiques "clés en main" à des établissements spécialisés de façon à profiter d'une rentabilité immédiate. C'est d'ailleurs en partie pour cette raison que les industries agroalimentaires entretiennent des relations privilégiées avec le secteur de la chimie et de l'équipement.

Les lancements de produits de grande consommation suivent des axes continus depuis plusieurs années, le goût, la qualité et la praticité **[SECO97]**. Les offres de goût se traduisent premièrement par un foisonnement des « saveurs », proposées pures ou mélangées.
Les progrès technologiques très importants réalisés par les industriels leur permettent de proposer des produits de très haute qualité de goût. Les succès des produits frais traiteur le démontrent. Enfin, la praticité à travers les emballages et les plats cuisinés reste une voie royale de développement.

Le lancement de nouveaux produits concerne dans la majorité des cas les grands groupes industriels (voir Tableau 8). Cela tient à une logique industrielle : au fur et à mesure que la qualité progresse, le coût d'amélioration de la qualité augmente. Paradoxalement, plus le produit est de consommation courante et de prix unitaire bas, plus il faut être gros pour innover car le progrès est fonction de l'effort consenti (lié à l'importance des équipes de recherche et à la qualité des chercheurs qui les constituent).

Tableau 8 : Nombre de nouveaux produits agroalimentaires par sociétés aux Etats-Unis en 1995 et 1996

Sociétés	1996	1995	1996/1995 en %
1. Philip Morris	191	169	+13%
2. Unilever	128	117	+9%
3. Nestlé	125	163	-23%
4. Grand Met	115	95	+21%
5. Sara Lee	115	176	-35%
6. CPC International	111	60	+85%
7. Campbell Soup	83	135	-39%
8. Wessanen USA	83	100	-17%
9. ConAgra	78	157	-50%
10. Hormel Foods	74	89	-17%
11. Nabisco Brands	68	109	-38%
12. Quaker Oats	55	59	-7%
13. General Mills	52	88	-41%
14. PepsiCo	47	32	+47%
15. H.J. Heinz	42	73	-42%
16. Borden	38	73	-48%
17. Dean Foods	38	44	-14%
18. M&M/Mars	32	67	-52%
19. Hershey Foods	28	19	+47%
20. Health Valley Foods	25	54	-54%
Total	1 528	1 879	-19%

Source: New Product News/PF.

Le goût

C’est l’un des axes les plus développés par les industriels de l’alimentaire car le goût est le principal attrait recherché par les consommateurs.
Aussi, voyons-nous se décliner les spécialités régionales, gastronomiques ou encore étrangères. Les échanges de population, les voyages ont fait découvrir de nouveaux goûts en provenance des pays du monde entier. Toutes les gammes, sur toutes les technologies s’enrichissent d’arômes indiens, chinois, antillais, japonais, mexicains, …
D’une façon générale, la qualité organoleptique des produits s’améliore, notamment grâce à des changements de technologie (exemple : les produits traiteur frais).

D'autre part, certaines marques cherchent à provoquer des ruptures de goût, soit en mélangeant des textures (exemple : produits bi-couches), soit en empruntant des goûts à d'autres univers (exemple : yaourt aux chamallow).
Enfin, les tendances de la démographie et des styles de vies ont engendré le développement de gammes de produits par exemple spécifiquement destinés aux enfants ou aux personnes âgées.

Les besoins des consommateurs ont évolué mais il ne s'agit pas, sous le prétexte de gagner des parts de marché, de sacrifier la réalité du goût. En effet, J.V PFIRSCH observe que « *Le "bon goût" se trouverait menacé par la "mondialisation". Le "goût" a sa "semaine", action associant les pouvoirs publics à des professionnels de métiers de bouche, permettant aux enfants des écoles de s'initier sous le regard intéressé des médias. Le "goût" a son Institut, l'Institut français du goût* (voir section 2.3.2)*, destiné à promouvoir recherches et cycles de formation en ce domaine. Les uns parlent de promotion des sens et d'épanouissement individuel, les autres de la nécessité de faire face à des menaces pesant sur la richesse des saveurs, la sauvegarde de la qualité de notre alimentation, la transmission des savoirs et des savoir-faire gastronomiques : le "goût" est dans bien des bouches, sur bien des langues. La notion de "goût", ses définitions légitimes et ses connotations symboliques constituent des enjeux économiques, sociaux et culturels importants. Peut être est-ce là un souci typiquement français*. »

La nutrition

La nutrition santé reste une tendance lourde qui correspond à un retour à une vie plus saine, à la recherche d'équilibre. Ce phénomène est largement favorisé par la vulgarisation de la médecine et par l'univers de plus en plus médicalisé dans lequel vivent les consommateurs (voir Tableau 9).
Quatre types de produits « santé » sont dénombrés :

- les produits de préventions : les « alicaments » ou tous les éléments qui soignent naturellement, grâce à leurs composants ou encore les produits biologiques qui garantissent une naturalité de leurs ingrédients
- les produits fonctionnels ou les produits enrichis et les compléments nutritionnels
- les produits énergétiques
- les produits diététiques

Mais plus qu'un marketing produit, la nutrition santé est surtout un axe de communication. EXEMPLE : Institut DANONE[6] **[DANO97]**.

[6] http://www.danone-institute.com/france/

Tableau 9 : Revendications nutritionnelles des produits agroalimentaires aux Etats-Unis de 1989 à 1996

Revendications	1989	1990	1991	1992	1993	1994	1995	1996
Réduction/allégé en calorie	962	1 165	1 214	1 130	609	575	1 161	776
Réduction/allégé en gras	626	1 024	1 198	1 257	847	1 439	1 914	2 076
Naturel	274	754	561	996	449	575	407	645
Réduction/allégé en sel	378	517	572	630	242	274	205	171
Sans additif/conservateur	196	371	526	631	543	251	167	143
Allégé/sans cholestérol	390	694	711	677	287	372	163	223
Ajout/riche en fibre	73	84	146	137	51	26	40	12
Réduction/allégé en sucre	188	331	458	692	473	301	422	373
Ajout/riche en calcium	27	20	15	41	14	23	21	35
Produits biologiques	140	324	370	510	385	446	538	645

Source: New Product News/PF.

L'emballage

Dans l'agroalimentaire, l'emballage fait partie intégrante du produit. Son rôle est essentiel, à en juger par la diversité de ses fonctions **[BURE89]**.
Il remplit avant tout, des fonctions techniques très importantes. Il représente une barrière entre la denrée alimentaire et le milieu environnant et, à ce titre, joue un rôle fondamental dans les processus de conservation et de distribution du produit. Il protège le produit alimentaire non seulement des contaminations microbiologiques en provenance du milieu extérieur, mais également contre l'action d'autres agents externes tels que l'oxygène ou la vapeur d'eau. Inversement, il doit préserver le produit contre les pertes d'arômes.
L'emballage intervient aussi très largement au niveau du marketing. Il permet de faire la promotion des produits, les positionne sur des segments de marché et des niveaux de gammes bien définis (voir section 2.3.5), offre des services au consommateur et véhicule de l'information de nature variée (voir section 2.1).
A l'heure où les préoccupations écologiques prennent de plus en plus d'importance, il doit en outre être facilement éliminé ou recyclé après l'usage.

Les conditionnements suivent la logique des marchés et surtout les attentes des consommateurs. L’emballage devient plus léger non seulement en raison de la législation mais aussi par économie.
Ils doivent garantir l’inviolabilité et être plus pratique et c’est là que réside l’une des évolutions majeures de ces dernières années. Les ouvertures deviennent plus faciles et refermables pour garantir la bonne conservation du produit.

2.3.5. La marque

Tous les jours, les consommateurs comparent les marques les unes avec les autres et les évaluent en fonction de leur goût et besoins.
La marque est un élément intrinsèque au produit et aussi un moyen de fidéliser le client **[KAPF89]**. Elle représente la mémoire du produit dans le sens où elle est le souvenir du consommateur une fois qu'il a utilisé le produit.
Elle permet enfin de maintenir le produit en permanence au faîte de sa mission : répondre mieux que ses concurrents à un type de besoin, à une certaine attente des consommateurs. L'amélioration des produits qui peut en résulter est l'un des éléments de la politique de qualité (voir section 2.2).
La marque est un moyen de concurrence certain pour les entreprises **[PLAN95]**. Plusieurs stratégies peuvent être envisagées, en voici quelques exemples.

Actuellement, les multinationales alimentaires procèdent à une restructuration complète de leurs portefeuilles de marques, dans tous les pays dans lesquels ils sont présents. Dans la plupart des cas, ils font disparaître des marques locales au profit d'une marque unique pour optimiser les budgets de communication. MARS fut l'un de pionniers il y a quelques années pour ses produits *Treets* et *Raider*. Tous les groupes ont désormais suivi.
D'autre part, les marques fortes ont pu être utilisées comme caution pour des développements sur de nouveaux univers. Par exemple *Nesquik*, marque de poudre chocolatée, qui sert de caution à des tablettes de chocolat, des bonbons, des briques, des crèmes desserts et des goûters au rayon frais.
Dans un autre cas de figure, pour supporter des produits touchant plusieurs univers de consommation, les marques existantes fortes sur leur domaine, s'adossent aux marques fortes de ces autres univers pour là aussi cautionner leurs produits.
Enfin, dans un ordre d'idées un peu différent, NESTLE a lancé une poudre chocolatée baptisée *Lion*, pour attirer les adolescents, également consommateurs de la barre chocolatée du même nom.

Rassurer les consommateurs sur la qualité des produits agroalimentaires ne suffit pas à les rendre fidèles dans leurs achats. En effet, ces derniers sont très friands de changement, de nouveauté. Leurs besoins changent avec leur style de vie et inversement. Face à ces exigences, l'entreprise agroalimentaire doit s'adapter.

Une autre démarche consiste à rendre le consommateur davantage actif et le prendre comme principal sujet d'étude. Son comportement servira de modèle pour le développement de nouveaux produits (Figure 2).

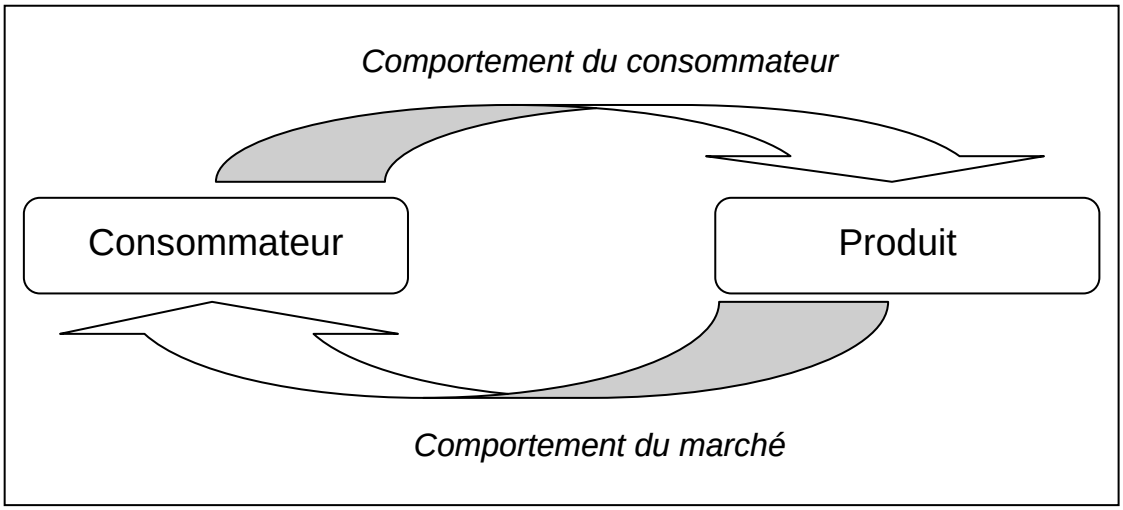


Figure 2 : Deux démarches dans la recherche d'innovation

3. Le comportement du consommateur dans le processus d'innovation

Le consommateur est le dernier maillon de la chaîne commerciale. Il est déterminant autant dans la réussite de lancement d'un nouveau produit que dans la fidélité de ses achats. Aussi, plusieurs disciplines se sont intéressées à lui, la plupart du temps pour connaître ses réactions vis à vis d'un produit.

3.1. Marketing

En marketing, tout commence et tout finit par le consommateur **[MARZ96]**. Pour connaître ses besoins et ses moyens, le contexte dans lequel il évolue ou pour chercher à le comprendre et mieux maîtriser ses réactions en ajustant l'action commerciale, le marketing fait appel à toutes les connaissances des sciences humaines fondamentales en les adaptant aux buts qu'il poursuit et aux méthodes qu'il met en œuvre.
La demande est représentée par la diversité des clients (clients potentiels, non clients, …). Pour raisonner et agir, les clients seront donc classés par groupes de comportement homogènes ou cibles d'actions particulières.
Le marketing-mix cherche à adapter le mieux possible la stratégie de marché en choisissant le bon produit pour le bon segment, communiquer la vraie image du produit avec les mots des clients, distribuer et rendre disponible au bon moment et au bon endroit.
Mais ces méthodes présentent des faiblesses. En effet, les échecs du marketing ont souvent été associés au manque de maîtrise du comportement du consommateur **[PETR88]**.

Par exemple, dans la plupart des cas, l'observation se limite à la mesure de réponses à court terme (enquêtes par sondage) alors qu'il est plus prudent de se focaliser sur du long terme (les résultats sont moins variables). Aussi plusieurs approches permettant de prévoir le comportement du consommateur à long terme ont été mises en place afin de tenter de dégager les voies à suivre en matière de produits et de services.

3.2. Economie

En économie, deux approches sont distinguées :

- la macro-économie (effet de masse) où l'étude de marché est liée à l'analyse de la demande (volonté des consommateurs à acquérir un bien) en cherchant à connaître les revenus et le prix pour lesquels la demande effective sera la plus forte.
- la micro-économie (effet individuel) où l'étude de marché cherche à connaître les moyens et les besoins des consommateurs pour contrôler sa demande.

Ceci est valable dans la mesure où l'individu se comporte de façon rationnelle, ce qui est loin d'être toujours le cas. En effet, de nombreux facteurs liés à l'économie comme par exemple les prix ou encore les crédits influencent le comportement du consommateur **[VANV94]**.

3.3. Psychologie

En psychologie, le comportement du consommateur sous la forme d'études de motivation sera analysé en fonction des phénomènes mentaux ou inconscients pour décrire les enchaînements d'événements internes à la personne. Ces derniers sont à l'origine de leur comportement et eux seuls peuvent expliquer que, dans les mêmes circonstances, deux personnes peuvent agir différemment **[HENA73]** et **[HENA79]**.
Malgré les apports bénéfiques de cette approche, plusieurs biais ont confronté les psychologues à son application au domaine du consommateur. Parmi eux, la difficulté à communiquer de façon simple et efficace leurs résultats aux gestionnaires d'entreprise a été déterminante **[PETR88]**.

3.4. Sociologie

La sociologie et la socio-psychologie ne s'intéressent au consommateur qu'en fonction de son appartenance à un contexte plus vaste : organisation, groupe social, société et culture **[VANV94]**.
Les études sociologiques mettent en évidence des régularités de comportement et permettent de fonder des hypothèses reliant identité et comportement.

3.5. Analyse sensorielle

En agroalimentaire, la recherche de séduction du consommateur est finalement assez difficile. Nous avons vu qu'il existe des moyens pour faire évoluer les produits en fonction des exigences des consommateurs. Mais comment être certain de pouvoir les satisfaire pour qu'ils renouvellent leur achat ?

Une des applications de l'analyse sensorielle, la cartographie des préférences (Figure 3) peut répondre à cette attente. Elle consiste à réaliser des tests de préférence consommateurs sur une gamme de produits cohérente. Ils sont affinés par une recherche de segments de consommateurs homogènes. Puis, la carte sensorielle est construite grâce à des dégustateurs experts. Enfin, les corrélations entre les données consommateurs et les données experts sont recherchées pour mettre en évidence les zones de préférences maximales et minimales **[SCHL92]**.

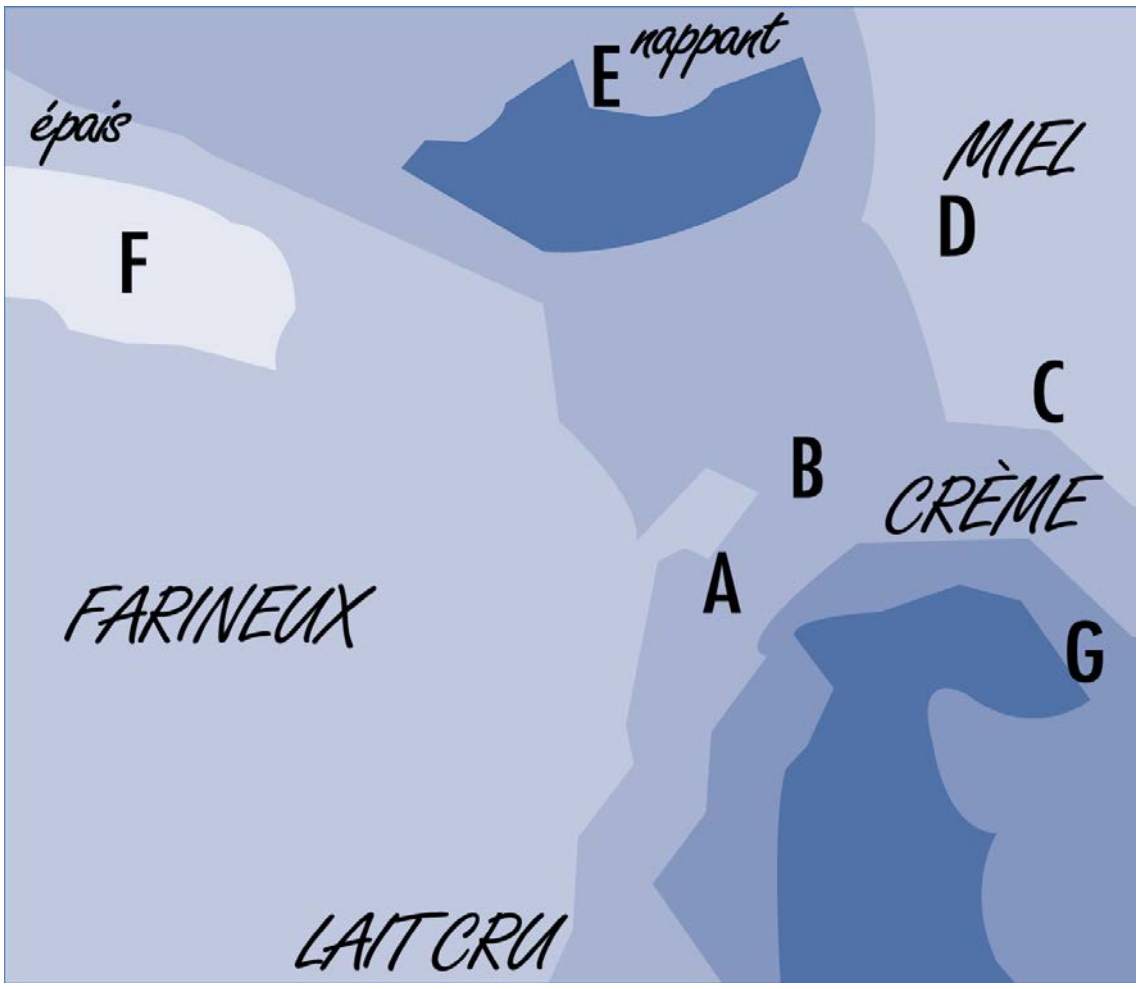


Figure 3 : Cartographie des préférences

La Figure 3 représente la position de sept produits laitiers désignés par les lettres A, B, C, D, E, F et G sur un plan où sont réparties les préférences des consommateurs. Les zones foncées correspondent aux meilleures notes et inversement les zones les plus claires montrent les notes les plus faibles.

Une telle cartographie montre que les consommateurs recherchent deux cibles de produits : les produits typés « crème et lait cru » et les produits « nappants ».

Cette technique demande au consommateur de s'exprimer quantitativement sur ses préférences à travers le plaisir que le produit lui a procuré.

Une autre idée très intéressante est de laisser formuler librement le consommateur afin qu'il décrive ses sensations. Ce type d'études est déjà employé dans les enquêtes par sondage (discussions de groupe ou entretiens individuels) mais le consommateur n'est pas en situation réelle de consommation. L'objectif est en effet d'obtenir une expression de ses motivations. Mais peu d'études ont cherché à obtenir des commentaires libres de consommateurs.

Dans la situation de concurrence actuelle, l'innovation devient de plus en plus nécessaire. Les consommateurs souhaitent et attendent des produits nouveaux et améliorés. L'innovation permanente semble être la seule façon d'éviter l'obsolescence de la gamme de produits d'une entreprise.
Dans le même temps il devient de plus en plus difficile de réussir le développement de nouveaux produits pour les raisons suivantes :

- **manque d'idées nouvelles**
- **morcellement des marchés, des volumes de ventes et profits réduits**
- **durcissement de la législation et prise de conscience des consommateurs**
- **augmentation rapide des coûts de développement des nouveaux produits**
- **taux d'échec élevé lors de la commercialisation**
- **diminution de la durée de vie des produits commercialisés avec succès**

Plusieurs solutions orchestrées soit par l'innovation, le marketing ou encore l'analyse sensorielle ont été proposées pour résoudre ce dilemme.
En parallèle, l'étude du comportement du consommateur permet de déterminer l'ensemble des relations complexes et les facteurs d'influences qui caractérisent la consommation.

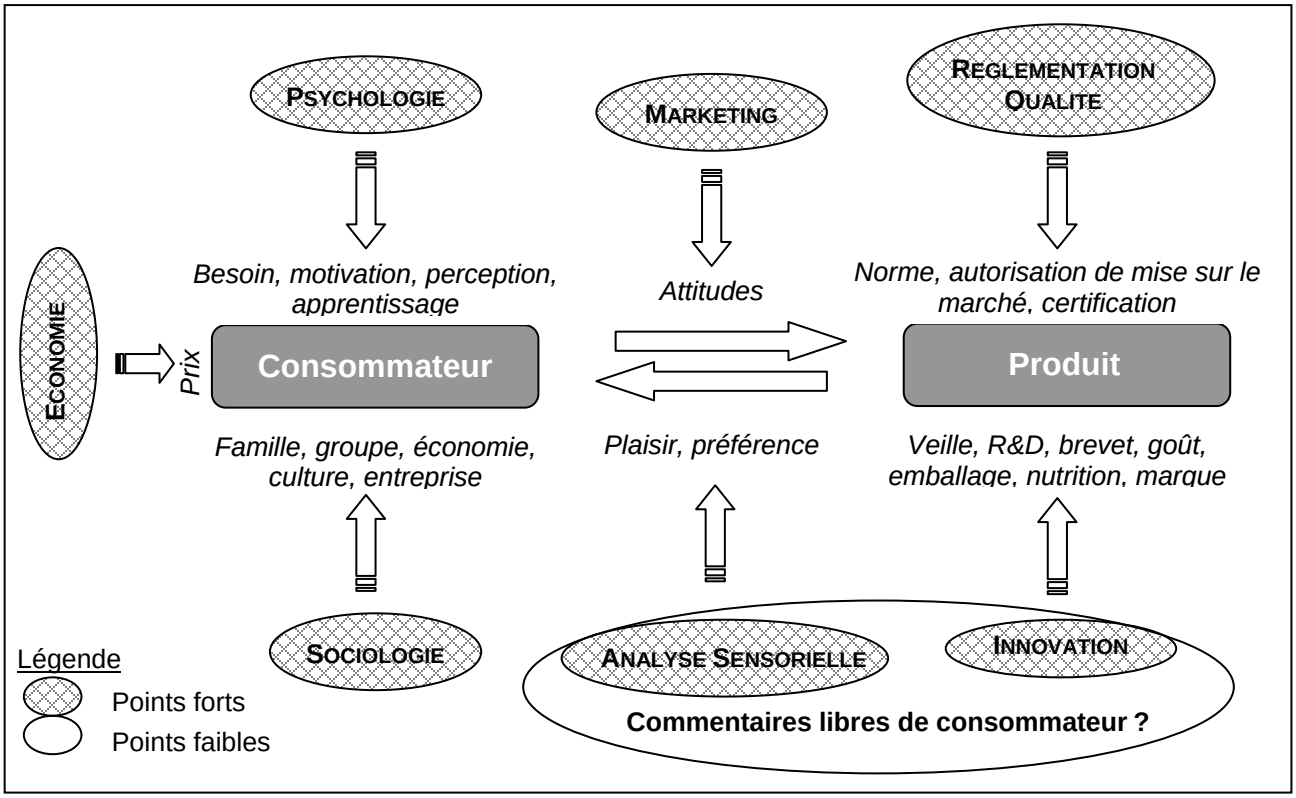


Figure 4 : Sept moyens de compétition en agroalimentaire

Si nous faisons le bilan des informations dont nous disposons sur l'interaction consommateur-produit, nous arrivons au constat suivant :

- **Au niveau du consommateur**
  - **l'approche économique fournie des informations sur le comportement des consommateurs à travers les prix des produits**
  - **la psychologie nous renseigne sur ses besoins, ses motivations envers les produits**
  - **la sociologie étudie les groupes de consommateurs en fonction des types de produits**

- **Au niveau du produit**
  - **les réglementations et la qualité contrôlent différents aspects du produit pour qu'ils soient acceptables pour le consommateur**
  - **l'innovation à travers la veille, la R&D, … propose des nouveaux produits pour conquérir de nouveaux consommateurs**

Enfin, le marketing et l'analyse sensorielle sondent le consommateur pour vérifier si le produit est bien perçu.

Chaque démarche est intéressante, complémentaire (Figure 4) et apporte de réelles solutions de compétition. D'une manière globale, ces solutions existantes sont satisfaisantes pour l'entreprise et c'est pour cette raison que nous les avons représentées en zone hachurée sur la figure. Par contre, aucune d'entre-elles ne s'intéresse vraiment à l'interaction directe entre le consommateur et le produit alimentaire au niveau de la perception sensorielle (zone sans hachure). C'est justement ce point faible que nous souhaitons aborder et développer dans ce projet de recherche.

L'étude de la compétition en agroalimentaire à travers l'innovation et le comportement du consommateur est de nature complexe. Il a semblé important au Groupe DANONE, commanditaire de cette étude, de s'intéresser à la perception sensorielle des consommateurs vis à vis des produits alimentaires car elle est une voie possible d'innovation qui n'avait jamais encore été abordée.

Ce projet a débuté en juin 1995 grâce à une collaboration entre le TEPRAL[7], et le Centre de Recherche Rétrospectives de Marseille[8] spécialisé en veille technologique et bibliométrie.
Nous espérons à travers cette expérience, obtenir un enrichissement mutuel vis à vis des deux types d'approches :

- **avec l'analyse textuelle sur le vocabulaire du consommateur**
- **avec la lexicométrie et la bibliométrie sur le traitement des données textuelles**

[7] Centre de Recherche de la branche Boissons du Groupe DANONE
[8] Université d'Aix-Marseille III

CHAPITRE II

Le succès d'un produit dépend aussi beaucoup du plaisir qu'il procure aux consommateurs et donc en grande partie de ses qualités gustatives.

Or, l'analyse sensorielle est spécialement destinée à examiner les propriétés organoleptiques d'un produit par les organes des sens **[AFN95a]***. C'est dans cet esprit que le projet de recherche sur la mise au point et la valorisation de nouvelles procédures pour mieux connaître la perception des produits alimentaires par les consommateurs s'est mis en place. L'expression spontanée du consommateur est intéressante pour l'analyse sensorielle pour comprendre comment il ressent le produit et surtout comment il l'exprime.*

Ce deuxième chapitre a pour objet de montrer comment les méthodes bibliométriques et lexicométriques nous ont aidés à travailler sur ces données, dans la mesure où elles ont apporté des solutions nouvelles à la problématique générale.

Aussi nous détaillerons la méthode d'analyse des commentaires libres que nous avons mise en place après avoir bien sûr caractérisé les données, les objectifs de l'étude et analysé les méthodes existantes à partir desquelles nous avons commencé à travailler.

CHAPITRE II : UNE METHODE D'ANALYSE DES COMMENTAIRES LIBRES DE CONSOMMATEURS

La manière dont les clients apprécient ses produits est vitale pour une entreprise qui travaille sur des marchés de consommation de masse. Mais il est difficile pour elle de savoir ce que le consommateur pense de ses produits : il n'existe pas un consommateur mais une multitude, tous d'avis différents, voire contradictoires.
De plus, le jugement "brut" des consommateurs est la plupart du temps de peu d'aide pour l'entreprise : très idiosyncrasique[9], souvent instable dans le temps, soumis à un nombre de facteurs d'influence considérable, souvent inconnus et non maîtrisables.
Enfin, les consommateurs s'expriment assez peu sur les produits achetés : les motifs de leur choix et leur appréciation des produits demeurent très souvent non verbalisés.

Et pourtant l'opinion du consommateur est indispensable : elle permet d'adapter les produits au client et donc d'assurer la bonne santé de l'entreprise. D'où la nécessité d'utiliser l'avis d'un ensemble de représentants des consommateurs.

Quelques expériences ont montré que les commentaires libres de dégustation issus des consommateurs constituent une source d'information riche et d'autant plus importante qu'ils n'apparaissent qu'en partie corrélés aux descriptions des experts sensoriels **[MART94]**. Autrement dit, le jury d'experts et le panel de consommateurs construisent chacun deux représentations du produit tout à fait indépendantes.

Il est donc important maintenant de montrer les particularités de l'information consommateur en analyse sensorielle.

1. Les caractéristiques des commentaires libres de consommateurs

Questionner un consommateur pour qu'il transcrive une appréciation sur le produit n'est pas une entreprise facile. Les techniques habituelles d'enquête via les questions fermées auront plutôt tendance à biaiser les résultats en :

- guidant les consommateurs dans leurs analyses
- provoquant des difficultés de compréhension du sens des descripteurs
- frustrant les consommateurs par des réponses non adaptées à leur perception
- créant des interférences sur les notations hédoniques

[9] Manière d'être particulière à chaque individu qui l'amène à avoir des réactions, des comportements qui lui sont propres.

Aussi, l'utilisation des questions ouvertes doit permettre au consommateur de participer à un test de manière directe, le plus proche possible des conditions normales de consommation et de ne pas avoir de contrainte de vocabulaire.

EXEMPLES : Commentaires libres collectés auprès d'un échantillon de consommateurs après la dégustation d'une bière (voir questionnaires en ANNEXES 2 ET 3)

Réponses à la question "*Quelles sont les sensations que vous avez ressenties en buvant cette bière ?"* :

→ Cette bière est aromatisée et rafraîchissante.
→ Un goût spécial, pas très marqué, mais qui se sent tout de suite.
→ Elle est agréable. Semble légère. Un peu amère. Elle désaltère.
→ Sensation de boire de l'eau. Pas de goût particulier. Elle est fraîche.
→ Rafraîchissante. Peu gazeuse.

Réponses à la question "*Citez les principales qualités de cette bière"* :

→ Sa couleur est parfaite. Son houblon est respectable.
→ Un peu de caractère, ce qui est une bonne bière.
→ Fraîche et légère. Semble peu alcoolisée.
→ Bulles fines. Facile à boire. Elle doit être facile à digérer.
→ Elle n'est pas forte. Elle se boit très vite. Elle ne reste pas au ventre. Elle n'est pas imposante à l'estomac.

Réponses à la question "*Dans quelles circonstances consommeriez-vous cette bière ?"* :

→ En dégustation ou en rafraîchissement.
→ En fin de soirée.
→ Après une journée de travail. L'été pour se délasser entre amis.
→ Pour se désaltérer.
→ Entre amis.

Mais les questions ouvertes introduisent malheureusement d'autres difficultés. Les réponses laissent transparaître tout d'abord une information à caractère flou lié au vocabulaire des consommateurs. Ces derniers emploient leurs mots avec leur signification dans leur environnement. Ceci entraîne obligatoirement une complexité d'interprétation pour ces données car le vocabulaire est hétérogène autant dans sa forme que dans son fond **[HOLL96]**.
Un deuxième problème important : ces données sont également complexes à traiter. Nous ne sommes plus dans le cas des réponses fermées ou semi-fermées qui représentent en statistiques, des données quantitatives.
Les réponses aux questions ouvertes sont assimilées à du texte libre avec tout ce que cela comporte comme difficulté à synthétiser.

1.1. Qualité des descriptions des consommateurs

De nombreux types de dégustations ou d'enquêtes sont entrepris en analyse sensorielle. Le plus souvent, ce sont des questions fermées qui sont utilisées.
Ces dernières sont plutôt utilisées dans des cas bien précis car ce type d'études connaît quelques désavantages **[JUAN86]**. En effet, les consommateurs ne sont pas toujours capables de répondre de façon pertinente à ce genre de questions, et ils peuvent d'autre part, être influencés par le contenu même des questions.
Par exemple, lorsque nous demandons à un consommateur de mettre une note de 0 à 7 sur l'intensité de l'odeur de citron dans une bière, il aura peut-être tendance à détecter effectivement cette odeur alors qu'autrement, il ne l'aurait probablement pas remarquée.

Nous pouvons également dire que le fait de devoir noter des critères à connotation négative ou positive, peut transformer son jugement de préférence. Par exemple, si nous lui demandons de noter une odeur de beurre, il peut se dire qu'il est anormal de trouver une telle odeur dans une bière et il sera donc peut-être plus sévère.

C'est justement dans l'optique de faire parler librement le consommateur et de lui rendre sa spontanéité que sont introduites les questions ouvertes dans les dégustations **[LEB93b]**. Les réponses ne sont plus restreintes à une grille de choix limitée. Les questions ouvertes permettent de demander aux enquêtés de répondre sous forme de phrases appelées encore commentaires.

1.1.1. Evaluation hédonique et évaluation descriptive

Plusieurs études ont montré que les consommateurs n'étaient pas performants pour décrire les produits d'une manière analytique. En effet, un grand nombre de facteurs intervient dans la dégustation des produits alimentaires (heure de dégustation, météo, région, publicité, …). Ils peuvent donc influencer plus ou moins la perception des consommateurs.
De plus, l'absence de langage commun introduit un problème de fiabilité pour ce type de données (voir section 1.1.5). C'est en partie pour ces raisons que l'intérêt des approches sur les consommateurs s'est porté uniquement sur des informations hédoniques via des questions sur les préférences **[ISSA92]**.

Or, il est possible d'introduire des questions ouvertes dans un questionnaire de préférence sans observer les biais liés aux questions fermées.
Le questionnement ouvert des consommateurs a une démarche différente de l'analyse sensorielle classique. Il s'agit plutôt d'une manière d'exprimer une sensation et non une description de leurs sensations.
Nous comprenons bien que le mode de questionnement soit primordial. Nous ne demanderons donc pas au consommateur de s'exprimer sur des caractéristiques ou des catégories sensorielles (voir section 4.1.3). Ces notions ayant des références différentes selon les individus, le contenu des réponses s'avérerait totalement incohérent.

D'autre part, l'ensemble des équipes d'analyse sensorielle du Groupe DANONE réalise un grand nombre de tests consommateurs pour collecter des informations hédoniques. L'insertion des questions ouvertes dans les questionnaires consommateurs aura donc tendance à produire une quantité importante d'information textuelle.

Plusieurs études consommateurs ont employé des questions ouvertes. Nous citerons en particulier deux études qui ont attiré notre attention par leur démarche et leurs résultats en l'analyse sensorielle :

- N. MARTIN montre que les consommateurs emploient fréquemment des critères de saveurs lors de dégustation de bières **[MART93]**
- le jury de consommateurs non avertis de G. TEIL utilise les caractéristiques *salé*, *acide* et *amer* pour l'arrière-goût et *onctueux*, *moelleux*, *fondant* et *crémeux* pour la texture pour décrire leur perception sur des fromages **[TEI94b]**.

A travers ces deux études nous avons pu mettre en évidence la richesse et la pertinence des commentaires des consommateurs.

1.1.2. Richesse de l'information

L'étude de G. TEIL sur la description des fromages par un jury de consommateurs non avertis a clairement démontré la richesse lexicale de leurs jugements **[TEI94a]**.
Mais avant de nous plonger dans le contenu lexical des commentaires libres, nous pouvons décrire l'image de sa structure lexicale. En effet, la distribution des fréquences de mots est considérée comme un indicateur pertinent pour la caractérisation des textes (voir section 1.2.1).
Le calcul d'un certain nombre de paramètres distributionnels d'un texte numérisé est réalisé à partir d'un tri à plat du lexique. Par exemple, dans le cas des réponses libres faites par des consommateurs après la dégustation d'un produit alimentaire, nous obtenons la répartition suivante :

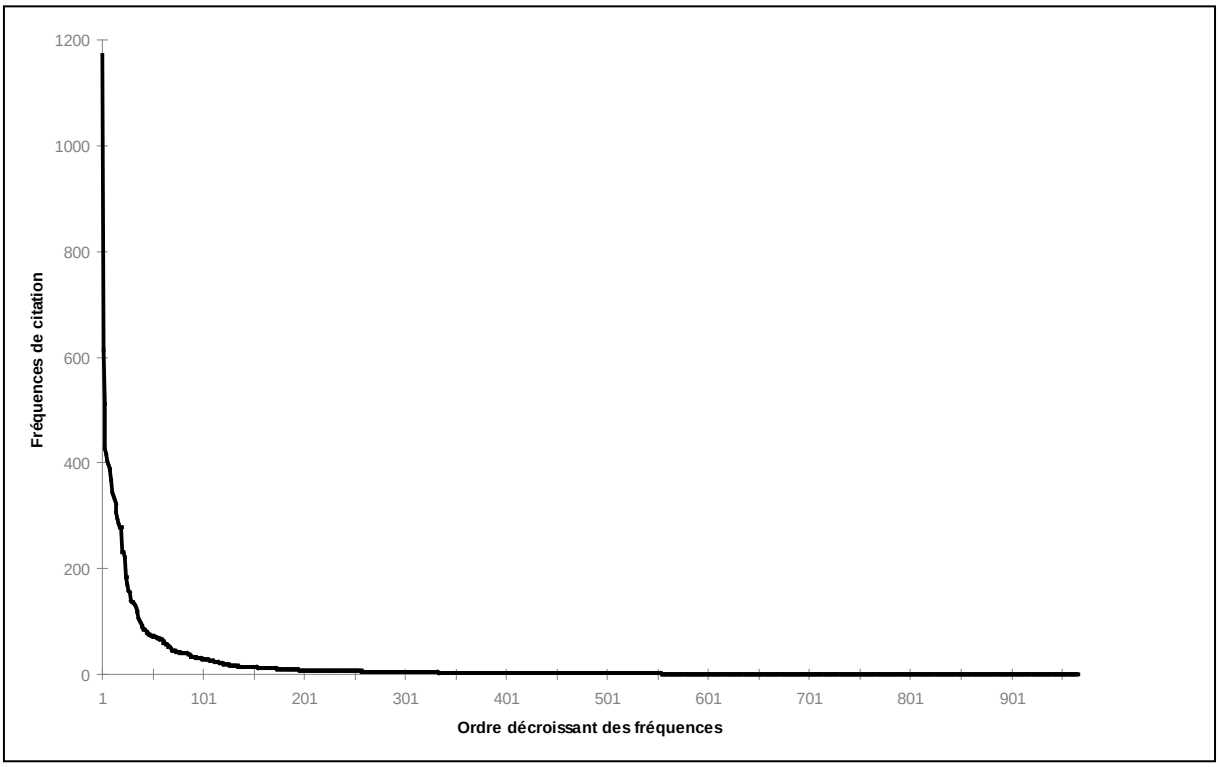


Figure 5 : Répartition du vocabulaire des commentaires libres

L'examen de la Figure 5 nous laisse pressentir le fait que cette distribution obéisse à une loi mathématique. En effet, la forme même de la courbe en hyperbole est traduite par une formule du type *Y = 1/X*.

ZIPF a justement été le premier à essayer de trouver une formule qui rende compte de l'échelonnement des fréquences à l'intérieur d'une distribution (et qui soit valable pour toutes les distributions constatées) **[ZIPF49]**.
La loi de Zipf illustre donc la relation schématisée sur la Figure 6 : le texte libre contient peu de mots à fréquence élevée mais beaucoup de mots à fréquence faible. Autrement dit, toutes les gammes de fréquences obtenues à partir de corpus de textes présentent des caractéristiques communes (voir la section 1.2.1 de ce chapitre).

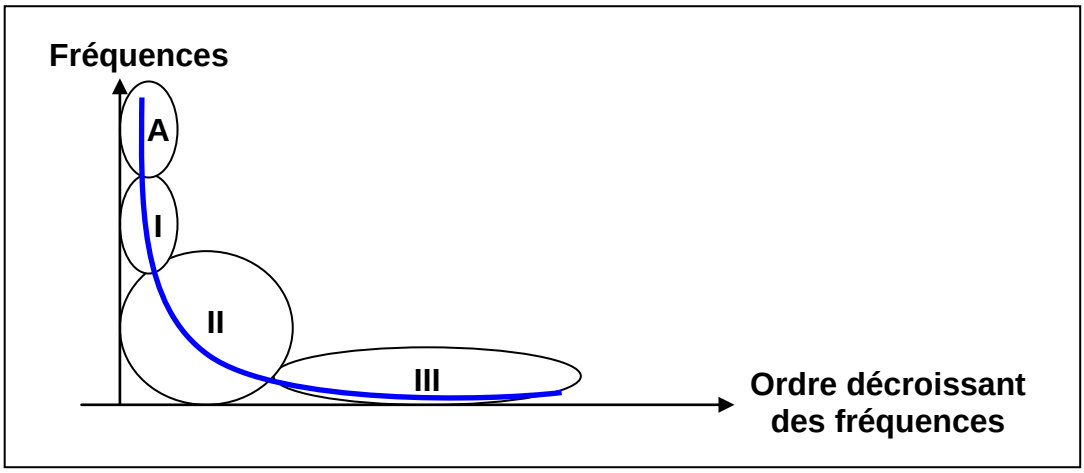


Figure 6 : Distribution d'un vocabulaire libre

Pour L. QUONIAM cette courbe modélisée sur la Figure 6 se segmente en quatre zones **[QUON92]** :

- la zone A représente les mots outils : ils sont indispensables dans une phrase mais inutiles dans un décompte lexical (les articles, les pronoms, …)

- la zone I englobe la partie triviale : elle contient les mots à très forte fréquence représentant les mots directement concernés par l'enquête (le nom du produit, les associations évidentes à un produit tel que *l'amertume* pour la bière, les verbes tels que *boire* ou *manger*,…)

- la zone II contient l'information intéressante : nous y trouverons le vocabulaire spontané du consommateur sur les sensations que lui procure la dégustation d'un produit alimentaire

- la zone III est associée au bruit : c'est un mélange de mots à très faible fréquence. Cet ensemble est complètement hétérogène puisqu'il peut à la fois contenir des mots outils, des mots triviaux ou des mots intéressants mais sous des formes différentes (orthographe, genre, nombre, conjugaison, … différents).

Il a souvent été démontré notamment en écologie **[LEGE84]** que ce genre de diversité de vocabulaire est un signe de maturité (voir la *Théorie de la communication de Shannon* à la section 1.2.1).

Pour les commentaires libres, cette richesse est en grande partie liée à la particularité d'une part, du questionnaire ouvert et d'autre part, du consommateur. Le fait de ne pas restreindre les appréciations entraîne la génération d'une multiplicité des termes. Une même sensation se traduit par différents descripteurs entre lesquels il existe généralement peu de concordances. La richesse potentielle des descriptions fournies par les consommateurs est de ce fait difficilement exploitable, d'une part, parce que les sujets ne possèdent pas de système de perception uniforme et, d'autre part car ils expriment différemment ce qu'ils ressentent **[ISHI87]**.

1.1.3. Caractéristiques d'énonciation

Les commentaires libres de consommateurs encore appelés réponses libres ou réponses aux questions ouvertes sont des données textuelles très particulières. En effet, L. LEBART et A. SALEM parlent de leurs caractères non seulement imprécis et multiforme mais aussi artificiel en termes de contexte, et relativement pauvre en termes de langue **[LEBA88]**. Pour eux, *le caractère oral des réponses donne des énoncés à la structure syntaxique approximative*.
L'observation des exemples à la section 1 confirme ces remarques. Nous pouvons dire qu'en règle générale, la grande majorité des commentaires comporte les caractéristiques suivantes :

- nombreuses énumérations
- nombreuses conjonctions de coordination
- phrases courtes, sans verbe
- phrases courtes avec des verbes passifs
- nombreux éléments de description associés à une expression de quantification

Au niveau du contenu, L. LEBART et A. SALEM **[LEBA88]** parlent du caractère artificiel des fréquences lexicales puisque selon eux, les enquêtés répondent tous à la même question. Cette particularité provoquerait une certaine répétition au sein du corpus créant de nombreuses tournures stéréotypées.
Les commentaires libres correspondent effectivement à une réponse à une même question mais dans un concept très large pour les consommateurs. Lorsque nous leur demandons "*Quelles sont les sensations que vous avez ressenties en buvant cette bière"* ils vont tenter d'exprimer leurs sensations avec non seulement leur vocabulaire mais aussi leur culture **[PFIR97]**.
Par contre, le fait qu'ils doivent s'exprimer sur un produit particulier va probablement engendrer quelques éléments communs de descriptions (par exemple l'amertume ou le côté rafraîchissant pour la bière).

1.1.4. Cohérence du vocabulaire

Quelques questions importantes ont préoccupé les études de G. TEIL **[TEI92a]** : *Est-ce que les consommateurs parviennent à s'exprimer seuls ? Ne collecte-t-on pas un ensemble d'impressions individuelles ?*

Lorsque les commentaires sont traités par des méthodes statistiques, G. TEIL retrouve l'existence d'un axe hédonique **[TEI94a]**. En effet, les consommateurs regroupent leurs descriptions autour des défauts et des qualités d'un produit.
Si les consommateurs ont du mal à trouver un consensus sur le vocabulaire d'évaluation sensorielle, ils n'ont aucun problème à se retrouver sur ce qui est bon ou mauvais pour eux.
D'autre part, elle montre également que le consommateur ne choisit pas ses mots au hasard. En corrélant les commentaires de consommateurs à ceux d'un jury d'expert, elle a pu mettre en évidence des thèmes du discours désignant des propriétés spécifiques des produits.

1.1.5. Difficulté d'interprétation

Pour un jury d'experts sensoriels, le vocabulaire de description est homogène car le but est de créer un consensus parfait au sein du jury. Aussi, les descripteurs doivent pouvoir être caractérisés de la sorte :

- pertinents : ils doivent être appropriés à la caractéristique sensorielle étudiée
  EXEMPLE : le caractère sucré est une saveur
- discriminants : ils doivent être susceptibles de varier d'une intensité faible à une intensité forte sur une même gamme de produits
  EXEMPLE : une forte odeur de céréales → une faible odeur de céréales
- monodimensionnels : ils ne doivent représenter qu'une seule dimension pour pouvoir être quantifiés
  EXEMPLE : un arôme de caramel
  CONTRE-EXEMPLE : un arôme fruité
- exhaustifs : ils doivent permettre de décrire le produit dans son ensemble
  CONTRE-EXEMPLE : une couleur jaune au-dessus et marron au-dessous
- non hédoniques : ils ne doivent pas avoir de connotations affectives.
  CONTRE-EXEMPLE : une saveur agréable

Le Tableau 10 montre un exemple de descripteurs utilisés par des experts sensoriels en bière pour analyser le produit sur ses propriétés organoleptiques [MEIL79].

Tableau 10 : Liste de descripteurs utilisés par les experts sensoriels en bière

Descripteurs	Composés de référence
Beurre rance	Diacétyle
Choux	Sulfure de diméthyle
Transpiration	Acide hexanoïque
Banane	Acétate d'isoamyle
Pomme verte	Acétaldéhyde
Rose	β-phényléthyléthanol
Sucré	Saccharose
Salé	Chlorure de sodium
Acide	Acide citrique
Amer	Caféine

Pour que ces descripteurs soient compris de la même façon par tous les juges, ces derniers sont soumis à un entraînement qui consiste à utiliser des produits représentatifs (composés de référence) et à respecter le même protocole de mesure.

Tableau 11 : Extrait du vocabulaire descriptif des commentaires libres de consommateurs

Termes affiliés à l'arôme et à l'odeur	Termes affiliés à la saveur	Termes affiliés à la texture
aromatique	Acide	alcool
aromatisé	Acidité	alcoolisation
aromatisée	Amer	alcoolisée
arôme	Amère	dur
avoine	Amertume	dure
caramel	Sucre	dureté
caramélisé	Sucré	eau
caramélisée	Sucrée	épaisse
citron	Vinaigré	épaisseur
fruité	Fade	coupé
fruitée	Fadeur	coupée
houblon	Douce	éventée
houblonnée	Douceur	gaz
houblons	âcre	gazeuse
malt	aigre	gazeux
médicamenteux	aigreur	fluide
métal	aigreurs	liquide
miel	âpre	lourd
rhum	âpreté	lourde
whisky	salée	lourdeur

La démarche est tout à fait différente avec les consommateurs. L'observation du Tableau 11 montre qu'en absence de consensus et de référence, les caractéristiques des descripteurs des experts énoncées plus haut ne sont pas retrouvées. Par exemple, le terme aromatique n'est ni monodimentionnel, ni unique puisque d'une part, il représente un ensemble de notions liées à l'odeur et à l'arôme et d'autre part, il existe plusieurs formes fléchies.

C'est en partie pour cette raison qu'il est difficile d'associer une caractéristique sensorielle à un descripteur cité par un consommateur. Nous employons ici le terme d'ambiguïté sensorielle (à distinguer de l'ambiguïté lexicale à la section 4.2.1) dans le vocabulaire des consommateurs. En effet, comment vérifier la pertinence d'un terme dans les commentaires libres ? Comment être sûr des caractères discriminants, monodimentionnels et exhaustifs de leur vocabulaire ? Comment éviter les connotations hédoniques puisqu'il s'agit de consommateurs ?

Dans un autre ordre d'idées, les techniques de traitement d'enquête à question ouverte proposent d'une manière générale deux modes de traitement :

- le postcodage qui consiste à codifier l'ensemble des réponses libres sous la forme d'une ou plusieurs modalités.
- l'analyse statistique des données textuelles par des méthodes factorielles.

Ces deux solutions apportent malheureusement autant d'inconvénients que d'avantages et ne résolvent pas le problème de la difficulté d'interprétation des commentaires libres de consommateurs. Les caractéristiques de ces deux approches seront traitées plus loin dans la section 2 de ce chapitre.

1.2. Valeurs statistiques des données textuelles

Les textes ont depuis longtemps fait l'objet d'études statistiques **[MUL92b]**. En effet, la statistique lexicale est l'étude quantitative de tous les mots d'un texte ou d'un corpus en fonction de leurs formes graphiques, leur appartenance à une catégorie grammaticale et de leur contenu sémantique.
La statistique textuelle a ouvert la voie de l'étude qualitative qui n'était pas envisageable avant d'avoir les moyens de calculs suffisants (outils informatiques) **[LEBA94]**.
Nous venons de voir que les commentaires libres de consommateurs sont des données textuelles particulières.
Avant de nous lancer dans leur analyse, nous allons exposer les caractéristiques de la statistique des données textuelles.

1.2.1. Lois statistiques

Les données textuelles libres constituent un objet relativement complexe pour les statisticiens. En effet, un texte est non seulement constitué de mots qui apparaissent avec une certaine fréquence et qui se succèdent suivant un certain ordre, mais il comporte également une dimension syntagmatique. Ceci sous-entend qu'un simple comptage des éléments constitutifs d'un texte ne suffit pas pour l'analyser. Nous ne pouvons pas négliger le sens qui se dégage des associations entre les mots (voir section 1.2.3).
Depuis les travaux de Zipf, de nombreuses études se sont succédées sur les statistiques appliquées aux textes. Malheureusement la complexité du matériau de base n'a pas encore permis de modéliser les textes d'une façon reproductible et fiable **[LEBA94]**.

Fréquences

En statistique textuelle, le choix des caractères délimiteurs (espaces entre les mots, points, virgules, …) permet de segmenter les commentaires en une suite de formes.
La suite de caractères non-délimiteurs bornée à ses deux extrémités par des caractères délimiteurs s'appelle une occurrence. Aussi, deux chaînes de caractères identiques représentent deux occurrences d'une même forme.
Le vocabulaire est représenté par l'ensemble de toutes les formes contenues dans le corpus de commentaires libres. Sa taille se mesure par le nombre d'occurrence. Elle est classiquement désignée de la façon suivante :

- la somme des effectifs correspondant à chacune des fréquences est égal au nombre des formes contenues dans le corpus. Nous mesurons alors le vocabulaire par la formule suivante :

$$V = \sum_{i=1}^{n} Vi^{*}$$

- la somme des produits (fréquences x effectifs) pour toutes les fréquences comprises entre 1 et Fmax, bornes incluses, est égale à la longueur du corpus. Nous mesurons donc la taille du corpus par la formule suivante :

$$T = \sum_{i=1}^{n} Vi \times i$$

[*] i = valeurs des fréquences

Loi de Zipf

Pour chercher à savoir *à quelle fréquence les mots apparaissent dans un texte littéraire,* Zipf a compté les occurrences des mots trouvés dans un corpus donné **[ZIPF49]**. Il les a ensuite classées par ordre décroissant de fréquence et il a affecté à chaque mot un rang, de 1 pour le mot le plus fréquent à n pour le moins fréquent.
En multipliant la valeur de chaque rang r par la valeur de la fréquence correspondante *f*, il a obtenu une constante C.
Il a appelé cette loi "*le principe du moindre effort*" puisque la probabilité d'occurrence d'un mot familier est bien plus élevée que celle des autres mots dans tous les types de textes et différentes langues **[BEAU94]**. En effet, ce type de distribution a été observé aussi bien dans les textes littéraires que dans les réponses à des questions ouvertes ou encore avec les références bibliographiques **[ROST93]**.
Il n'a pas proposé de représentation graphique de sa loi, mais en restant dans le même esprit que les précédentes lois, il est facile d'imaginer que la loi de Zipf s'applique parfaitement au modèle de distribution du vocabulaire libre (voir section 1.1.2).

Cette loi a fait l'objet de nombreuses recherches avec des normes de dépouillements différentes : sur des classes de fréquences, sur les rangs de fréquences ou encore sur les fréquences cumulées **[BEAU94]**.
Mais la formulation mathématique ne permet pas encore de déterminer des indicateurs pour connaître le poids statistique d'un ensemble de formes suivant sa répartition dans le vocabulaire d'un corpus. Il semble en effet qu'il y ait un nombre élevé de paramètres entrant en jeu, qui ne sont pas toujours simples à évaluer et encore moins à interpréter.
La loi de Zipf sera peut être employée à sélectionner les mots les plus représentatifs des textes initiaux en découpant le modèle de distribution vu à la section 1.1.2 en zones et leur donner un sens statistique **[ROST96]**.

Théorie de la communication Shannon

Cette théorie appliquée dans de nombreux domaines, propose des critères de mesure pour caractériser une distribution de données.
Nous avons déjà mentionné les travaux d'évaluation de la richesse d'un milieu en écologie à la page 9 **[LEGE84]**. Cette richesse est liée au nombre d'espèces différentes dans le sens où plus il y a d'espèce, plus l'entropie augmente et plus l'écosystème s'équilibre. Nous retrouverons le même genre d'observation en génétique où la vigueur d'une espèce croît de façon inverse à la consanguinité.
Pour son application en sciences de l'information et de la communication, des auteurs ont cherché à l'utiliser dans le cadre d'une mesure synthétique d'une distribution bilbiométrique **[LAFO92]**.

L'ensemble de ces lois statistiques nous paraît bien théorique et peu applicable à nos corpus. Cependant, selon H. ROSTAING : *La connaissance de ces lois reste indispensable à la réalisation d'un traitement statistique des données textuelles. La caractérisation hyperbolique de ces distributions est une notion fondamentale. Toutes les méthodes statistiques ne sont pas bonnes à employer car elles sont bien souvent construites sur le principe d'une répartition normale. L'emploi de la moyenne en est un très bon exemple. La valeur moyenne de la fréquence des mots dans un texte n'a pas*

beaucoup de sens. Actuellement, la seule comparaison possible est celle qui oblige à évaluer la différence entre les deux distributions complètes. C'est pourquoi les méthodes d'analyses des données sont purement descriptives. Elles ne cherchent pas à découvrir des modèles régis par des lois de distributions Gaussiennes **[ROST96]***.*

Richesse lexicale

La statistique lexicale, encore appelée lexicométrie, nous l'avons vu est l'étude de l'organisation du vocabulaire dans le discours ou autres domaines littéraires (voir section 1.2).
C'est dans cette optique que **[MUL92a]** a proposé de mesurer la richesse lexicale (ou richesse du vocabulaire) à partir d'une comparaison de la distribution du vocabulaire du corpus et d'une courbe théorique. Elle est surtout utilisée dans le cadre d'études stylométriques pour déterminer l'identité d'un auteur, la date ou l'époque de l'écriture d'un manuscrit.

Soit, V' le vocabulaire attendu dans un corpus de taille T[10], alors :

$$V'(\underset{T}{T}) = V - \sum_{i=1}^{n} Vi \times (1 - \underset{T}{T'})^{i}$$

Cette formule est seulement valable dans l'intervalle suivant : Fn < T' < (T-Fn) avec Fn comme fréquence maximale.

La richesse lexicale a été en réalité un des grands thèmes de recherche en statistique lexicale **[BERN88]**. Malheureusement, aucun indicateur fiable a pu être actuellement mis en évidence.

1.2.2. Chaîne de traitement

La façon d'aborder l'étude des données textuelles s'inscrit dans un processus classique de traitement statistique. Ce dernier se présente sous la forme d'une chaîne de traitement qui consiste à diviser le travail en plusieurs phases :

- la collecte d'information
- le traitement linguistique ou lexical du vocabulaire
- le traitement statistique du vocabulaire
- la représentation graphique du traitement statistique
- l'interprétation de la représentation graphique

Chacune d'entre elles devra faire l'objet d'une attention particulière en lien avec les préoccupations issues du domaine d'application. D'autre part, elles se succéderont dans l'ordre énoncé ci-dessus car elles sont fortement dépendantes les unes des autres.

[10] la valeur de N' doit être inférieure à la taille du corpus à mesurer

Il s'agit d'un critère important pour des applications informatiques car il est nécessaire qu'elles communiquent entre elles (par exemple pour l'importation ou pour l'exportation des données).

1.2.3. Notion de méta-information

L'ensemble de l'information dont nous disposons pour effectuer des analyses mais qui n'est pas utilisé est désigné sous le terme de méta-information **[LEBA94]**.
Dans le cas des données textuelles, elle est particulièrement abondante. En effet, chaque mot peut être assimilé à une définition ou encore à des règles de grammaire lorsqu'il est employé dans une phrase.
Il va donc être important de déterminer quel niveau de pertinence nous voulons choisir pour respecter ces différents niveaux de méta-information.

A titre d'exemple, si en recherche documentaire nous désirons travailler sur des mots clés (variables qualitatives de présence-absence), nous pouvons nous limiter à la construction de matrices classiques. Ces dernières serviront ensuite de base pour les calculs statistiques (analyses factorielles, classification, …). Les documents ne seront plus des textes au sens linguistique du terme mais des listes de mots.
A ce niveau, il faut prendre en considération l'importance de la segmentation du texte. En effet, au cours du traitement statistique, les mots qui étaient à l'origine liés par des règles grammaticales, sémantiques et pragmatiques ont été isolés. Nous nous retrouvons alors devant le constat paradoxal suivant : comment obtenir un gain en signification statistique sans déclencher une perte d'information linguistique ?
Ceci engendre quelques problèmes d'ambiguïté lexicale souvent dus aux polysémies et homographies.

Cette partie nous a montré que les commentaires libres de consommateurs peuvent être qualifiés de données textuelles au même titre que les entretiens, les discours ou tout textes littéraires.
Mais ils ont bien entendu leurs spécificités qui les rend difficiles à traiter.
Si cette approche est nouvelle pour l'Analyse Sensorielle, elle est depuis longtemps abordée dans les disciplines littéraires. L'analyse des méthodes existantes en traitement des données textuelles peut nous aider à aborder nos travaux sous un nouvel angle.

2. Analyse des méthodes existantes en traitement des données textuelles

L'information textuelle prend de plus en plus d'importance dans notre société. En effet, nous remarquons depuis une vingtaine d'années que le volume de connaissances que nous ingurgitons augmente sans cesse. Deux phénomènes sont fédérateurs de cette constatation : la communication est davantage écrite qu'orale et les moyens de diffusion sont plus faciles, rapides et fiables.

Aussi l'information textuelle est-elle présente partout et finalement plusieurs domaines scientifiques et littéraires sont amenés à la manipuler. Chaque démarche part bien sûr d'objectifs particuliers liés au domaine en question. Pourtant, elle est bien souvent abordée grâce à l'interconnexion de plusieurs disciplines (linguistique, informatique, statistique, sociologique, scientométrique, aide à la décision : text mining…).

Il était donc intéressant pour nous, d'étudier d'une façon générale, les différentes approches du traitement des données textuelles pour, d'une part, se familiariser avec le matériau et, d'autre part, positionner notre démarche.

2.1. Dans la bibliographie

Trois axes principaux ont guidé notre recherche sur l'existant en traitement des données textuelles :

- la collecte de l'information textuelle
- le codage de l'information textuelle
- le traitement statistique et la représentation graphique de l'information textuelle

2.1.1. Sur la collecte

La conception, la réalisation et le traitement des enquêtes sont de très bons exemples de coopération interdisciplinaire et interprofessionnelle dans le domaine de l'acquisition des connaissances.

Les techniques d'enquêtes

De nombreux travaux sur le mode de questionnement ont été menés sur la qualité de l'information dans les enquêtes (**[ASU92], [GRAN93], [YVON90], [LION91]**).
En effet, la qualité des données collectées est fonction d'un grand nombre de facteurs tels que la conception du plan de sondage, la cohérence et la clarté du questionnaire, les contrôles de terrains, les modes d'interrogation et les dispositifs techniques correspondants (matériels et logiciels), la formation et la rémunération des enquêteurs, les nombreux facteurs qui conditionnent l'entrevue, la saisie, les redressements et prétraitements, les tests de cohérence. Souvent il faut ajouter à cela des contraintes temporelles, financières et juridiques.

Nous retiendrons parmi tous ces paramètres les idées principales suivantes :

- la constitution de l'échantillon de consommateur est le plus souvent effectuée par des méthodes non aléatoires par exemple, en déterminant des quotas en fonction des distributions connues de la population **[DEVI92]**. En effet, il est difficile de déterminer un échantillon représentatif surtout en analyse sensorielle à cause des différences de récepteurs sensoriels (autant en quantité qu'en qualité) et psycho-sociologiques.
- l'échantillon peut également être extrait d'un panel **[ASU91]** de consommateurs recrutés par un institut de sondage (IPSOS, …). Les panels peuvent être spécialisés suivant des caractéristiques ou des thèmes particuliers. C'est un moyen sûr, simple et rapide d'obtenir un échantillon de consommateurs.
- les libellés des questions jouent un rôle fondamental. En effet, **[GREM87]** a observé qu'il est souvent difficile de trouver deux libellés distincts, pour deux questions fermées dont les contenus sont similaires, donnant les mêmes résultats en termes de pourcentage.

Ces réponses sont d'autant plus variables qu'elles changent de place à l'intérieur du questionnaire ou qu'elles sont lues par des personnes d'origine différente ou encore que la longueur du libellé est plus ou moins long **[GREM92]**.

Les questions ouvertes

L. LEBART et A. SALEM citent quatre cas d'utilisation des questions ouvertes **[LEBA94]** :

- *Pour économiser le temps d'interview*
  Bien que les réponses libres et les réponses guidées fournissent des informations de nature différente, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue et de tension.

- *Pour expliciter les réponses à des questions fermées*
  C'est la question complémentaire classique : "pourquoi ?". Les explications concernant une réponse déjà donnée doivent nécessairement être fournies de façon spontanée. Une batterie d'items risquerait de proposer de nouveaux arguments qui ne pourraient qu'entacher l'authenticité ou la sincérité de l'explication.

- *Pour critiquer et évaluer la qualité de l'information*
  "Vous venez d'être interrogés longuement sur vos conditions de vie, y a-t-il des sujets importants que vous auriez aimé voir aborder ? Avez-vous des remarques à formuler ?"
  Les questions de ce type peuvent dans certains cas remettre en cause d'importantes parties du questionnaire, mettre en évidence ses a priori et ses lacunes.

- *Pour recueillir une information spontanée par nature*
  Les questionnaires des enquêtes de marketing abondent en questions de ce type. Citons par exemple : "Qu'avez-vous retenu de ce spot publicitaire ?", "Que pensez-vous de cette voiture ?".

L. LEBART rajoute que *les questions ouvertes sont intéressantes pour des questions portant sur des attitudes, besoins, motifs dont les contours sont a priori mal connus, imprécis et difficilement catégorisables* **[LEB93b]**.

Souvent, les questions ouvertes sont utilisées dans la phase de préparation du questionnaire. Elles permettent de mettre au point les modalités des questions fermées. Les questions ouvertes peuvent être également mélangées avec des questions fermées dans certains questionnaires de façon à éviter le plus possible de distorsions **[JUAN86]**.

Les impacts du mode de questionnement ouvert des consommateurs en analyse sensorielle n'ont, à notre connaissance, jamais été évoqués dans la littérature. En effet, G. TEIL **[TEI94a]** n'utilisait pas les questions ouvertes et les études du CREDOC[11] ne relatent pas un travail d'analyse sensorielle (**[AUCO91]**, **[BEA93a]**, **[BEA93b]**, **[BEAU94]**, **[BEAU95]**, **[LAHL92]**, **[LAHL93]**, **[LION91]**, **[YVON90]**). Enfin, N. MARTIN dans sa thèse introduit le simple libellé suivant : *"Décrivez l'ensemble des sensations que vous a procuré ce produit ?"* **[MART93]**.

D'autre part, l'étude de V. BEAUDOUIN et S. LAHLOU montre que le mode auto-administré sous forme écrite est davantage adapté pour recueillir des réponses longues, riches en vocabulaire, syntaxiquement bien formée **[LAHL93]**. Cette constatation particulièrement intéressante sera mise en pratique à la section 4.1.2.

2.1.2. Sur le codage

Deux approches nous intéressent particulièrement en ce qui concerne la préparation des unités textuelles dans la perspective d'effectuer un traitement statistique. Il s'agit en effet dans un premier temps, de réduire la dispersion du vocabulaire pour diminuer la taille du vocabulaire d'une part et augmenter les fréquences de citation d'autre part. Dans un deuxième temps, nous devons garantir une perte minimale d'information.

Homogénéisation du vocabulaire

Pour effectuer des décomptes d'unités ou encore les additionner entre elles, ce qui se résume à effectuer des calculs statistiques, le vocabulaire d'un corpus doit être dans un premier temps segmenté en unités minimales[12] (voir section 1.2.1). Cette procédure permet de manipuler des éléments de la même importance pour les comparer. Dans la pratique, l'application de ces principes généraux implique que soit définie une norme permettant d'isoler de la chaîne textuelle les différentes unités sur lesquelles porteront les dénombrements à venir **[LEBA94]**.
La seconde opération consiste à réduire la variabilité des unités minimales. Pour procéder à une homogénéisation du vocabulaire, il existe des méthodes diverses. Les différents travaux sur l'exploitation des données textuelles montrent deux principaux courants qui découlent de deux modes de travail, l'un manuel et l'autre automatique.

[11] Centre de Recherche pour l'Etude et l'observation des conditions de vie
[12] Unités que l'on ne décomposera pas plus avant

Méthodes manuelles

L'analyse de contenu ou post-codage pour L. LEBART et A. SALEM, consiste à traduire les réponses des questions ouvertes à travers une grille de lecture afin de les transformer en une ou plusieurs questions fermées a posteriori **[LEBA94]**. Pour réaliser ceci, il aura fallu au préalable effectuer une première analyse intellectuelle de toutes les réponses pour prendre connaissance des éléments qui seront codés. C'est suivant cette procédure que le dépouillement des questions ouvertes est classiquement abordé en marketing.
Le même principe de travail se retrouve en indexation ou classification documentaire (**[COUR76]** et **[LECR90]**).
Cette méthode pose le problème du degré de finesse que nous pouvons atteindre. En effet, s'il y a plusieurs aspects dans les réponses, nous obtenons plusieurs items possibles. Ceci sous-entend qu'il faudra distinguer plus ou moins de modalités pour traduire les nuances avec plus ou moins de précision.
D'autre part, la présence de modalités telles que "autres" ou "divers" attire un grand nombre de thématiques périphériques.
D'autres contraintes comme le coût élevé lié au facteur temps ainsi que la subjectivité liée à la personne qui analyse le document ont poussé les chercheurs à trouver une solution automatique.

Méthodes automatiques

Elles consistent à réduire le vocabulaire d'un corpus de réponses libres grâce à des techniques computationnelles. Elles découlent de plusieurs disciplines qui ont toutes un point commun : le texte. Mais elles se distinguent tout de même par le fait que leurs buts soient différents.

Un ordinateur peut extraire assez facilement les mots d'un texte, le problème revient à sélectionner les termes les plus représentatifs. La première étape consiste à éliminer les mots vides qui sont communs à toutes les langues. Approximativement la moitié des termes utilisés dans un texte sont des termes grammaticaux qui ont un sens seulement dans le contexte de la phrase où ils sont utilisés. Par exemple les articles, conjonctions, prépositions, pronoms, adjectifs numéraires etc … sont des termes grammaticaux. Il est possible d'en dresser la liste composée de quelques centaines d'éléments, et à partir de cette dernière, de les éliminer automatiquement des textes originaux. Les autres éléments de la phrase sont des termes lexicaux qui ont un sens indépendant de leur contexte d'utilisation. Toutefois cette sélection rudimentaire révèle des défauts évidents :

- tous les termes lexicaux sont retenus, qu'ils soient ou non représentatifs des sujets traités dans le texte.
- les mots sélectionnés ne sont pas normalisés, par exemple un même verbe est alors sélectionné plusieurs fois, correspondant aux différentes formes grammaticales utilisées dans le texte.
- Seuls les uni termes sont pris en compte, délaissant les mots composés, par exemple "bière sans alcool" ou "arrière goût", qui se révèlent être souvent plus représentatifs et moins ambiguës.
- Les difficultés inhérentes au traitement du langage naturel, c'est-à-dire les problèmes de polysémie et de synonymie, ne sont pas résolues.

Les travaux que nous allons exposer tentent de résoudre ces problèmes de façon automatique.

Seuil sur la fréquence et la taille des mots

L. LEBART réduit la dispersion du vocabulaire en supprimant l'ensemble des mots dont la fréquence d'apparition dans le corpus est inférieure à une fréquence limite **[LEBA94]**. De la même façon, N. MARTIN et M. ROGEAUX retiennent les mots cités pour un même produit par au moins 10% des consommateurs **[MART94]**.
Une autre technique consiste à éliminer tous les mots contenant moins de quatre lettres car pour L. LEBART, ces derniers correspondent souvent à la plupart des mots outils (le, la , de, des, un,…). En effet, selon L. LEBART et A. SALEM, la réalisation d'une telle analyse n'a de sens, d'un point de vue statistique, que si les termes apparaissent avec une certaine fréquence **[LEBA88]**. Ceci est caractéristique des techniques d'analyse d'inertie. Les mots à faible fréquence ont forcément un profil très marginal et donc créent une distorsion du nuage de points.

Cette pratique lexicométrique employée dans le but de réduire la dispersion du vocabulaire n'est pas acceptable. En effet, il s'agit d'une pratique trop radicale et trop simpliste qui consiste à résoudre un problème en créant plusieurs autres. Comment être certain d'avoir éliminé les termes qui ne nous intéressaient pas et gardé les termes importants de cette manière là ? Dans nos commentaires libres, parmi les mots de moins de quatre lettres, nous trouvons des formes que nous souhaitons garder (gaz, dur, eau, fin, sec, …). Inversement, nous pouvons également trouver des mots outils de plus de trois lettres (quand, autre, après, presque, …). Enfin, si certaines formes apparaissent avec une fréquence faible ceci ne signifie par forcément qu'elles ne seront pas importantes au point de les éliminer !

Tableau 12 : Extrait d'un vocabulaire de commentaires libres, fortes et faibles fréquences

Formes	Frequences
goût	459
de	343
bière	325
pas	288
elle	273
la	244
….	….
autre	26
…	…
acidité	1
âpreté	1
pâteux	1
pâle	1
pétillant	1
éventée	1
whisky	1
volupté	1
vinaigré	1
néanmoins	1
…	…

Le Tableau 12 nous montre d'une part qu'il existe des mots outils de plus de trois lettres et de fréquence supérieure à 1 à l'intérieur les vocabulaires de commentaires libres de consommateurs et d'autre part que parmi les faibles fréquences nous retrouvons des formes descriptives importantes.

Lemmatisation

D'une façon générale, la lemmatisation consiste à regrouper sous une forme canonique[13] l'ensemble des formes fléchies d'une même unité minimale. Habituellement, cette opération est réalisée suivant une norme lexicologique appropriée au contexte général du texte **[MUL92a]**.
EXEMPLE :

- regroupement des formes fléchies des verbes sous leur forme infinitive :
  mangé, mangés, mangée … → manger
- regroupement des formes fléchies des substantifs sous leur forme au singulier :
  arômes, odeurs, goûts→ arôme, odeur, goût
- regroupement des formes fléchies des adjectifs sous leur forme au masculin singulier :
  aromatisés, aromatisées, aromatisée → aromatisé
- regroupement des formes fléchies élidées sous leur forme sans élision :
  l'arôme, l'arrière-goût → arôme, arrière-goût

La lemmatisation apporte l'avantage d'augmenter la fréquence des formes citées dans les textes mais ceci engendre l'inconvénient de regrouper des formes qui n'ont pas le même contexte de citation. C'est pour cette raison qu'il existe deux courants de pensée : les adeptes de la lemmatisation qui privilégient l'étude quantitative du texte et les adeptes de la non-lemmatisation qui privilégient l'étude qualitative. Ces derniers auront recourt à d'autres techniques telles que l'analyse des concordances ou les index (voir plus loin).

Tout de même, la lemmatisation permet non seulement d'obtenir une grille de lecture unifiée mais aussi de réduire considérablement la taille du vocabulaire. Cette dernière particularité résout le problème des matrices creuses[14] qui est classique en l'analyse des données textuelles. Selon S. LION : *en réduisant le nombre de formes graphiques par agrégation de certaines d'entre elles, elle diminue le nombre de cases vides et rend plus robuste une analyse en terme de profils lexicaux. En effet, sur des tableaux statistiques trop clairsemés, comme c'est souvent le cas, les profils lexicaux des individus ont tendance à être trop semblables. La multiplication des cases vides "écrase" les profils des individus, et multiplie d'autant l'impact des artefacts sur l'analyse finale* **[LION91]**.

Cependant, une lemmatisation automatique parfaite n'est pas possible sur un corpus de texte sans analyse syntaxique globale de la phrase à cause de nombreuses ambiguïtés lexicales.

[13] Forme naturelle, intrinsèque, principale.
[14] Matrices contenants de nombreux zéros

Analyse morpho-syntaxique

Elle permet d'affecter une catégorie grammaticale accompagnée d'un lemme de rattachement à chacune des formes graphiques d'un texte.
EXEMPLE :

Les [[DETDEF LE]] problèmes [[NOMMP PROBLEME]] matériels [[ADJMP MATERIEL]] , [[PONCT-FAIBLE VIRGULE]] une [[DETINDF UN]] certaine [[ADJFS CERTAIN]] angoisse [[NOMFS ANGOISSE]] vis-à-vis-de [[PREP VIS-A-VIS-DE]] l [[DETDEF LE]] avenir [[NOMMS AVENIR]] . [[PONCT FORTE POINT]]

Les problèmes d'ambiguïté lexicale sont ainsi écartés puisque l'information grammaticale qui était perdu lors de la lemmatisation est ici associée à la forme.

Cette méthode est intéressante car elle nous permet de retrouver une partie de la méta-information. Elle est malheureusement dépendante de dictionnaires qu'il faudra sans cesse agrémenter et incrémenter lorsqu'une nouvelle forme graphique apparaîtra. Enfin, elle ne résout pas complètement les problèmes d'ambiguïtés sémantiques et pragmatiques.

Traitement du langage naturel

Les outils qui proposent ce genre de traitement sont tous issus de la même idée de base : apprendre à la machine à fonctionner comme un cerveau humain.

Pour modéliser la langue, les linguistes ont établi des théories complexes telles que la théorie de la grammaire syntagmatique généralisée ou GPSG **[GAZD95]**, la théorie de la grammaire syntagmatique guidée par les têtes ou HPSG **[POLL94]** et la théorie de la grammaire lexicalisée d'arbres adjoints ou TAG **[EJER95]**.
Ces théories veulent être confrontées aux données empiriques dans un but de falsification. Aussi les statisticiens et les informaticiens ont apporté de nouvelles solutions (probabilités et puissance de calcul) pour appliquer les théories linguistiques au travers d'outils de traitement du langage naturel ou de représentation des connaissances.
Ce domaine est sensible pour les industriels car ils sont de plus en plus confrontés à des documents textuels en grand nombre. Mais, il est également émergent. C'est pour cette raison qu'il s'agit de projets à long terme, souvent subventionnés car ils restent très coûteux[15].

En effet, nous comprenons facilement que pour réaliser une application de traitement automatique du langage naturel non limitée à un domaine particulier, nous aurons besoin d'une quantité considérable de connaissances lexicales, syntaxiques et sémantiques.

[15] EXEMPLE : le Projet Eurêka GRAAL ou Grammaires Réutilisables pour l'Analyse Automatique des Langues : 23 Mecu, 150H.années conduit par un consortium international réunissant GSI-ERLY, EDF, Aérospatiale, Renault, Xerox (en France), ISSCO (en Suisse), ILSP (en Grèce), Lingsoft, Nokia (en Finland), Centro Ricerche FIAT, IRST (en Italie), ILTEC (au Portugal).

Enfin, ces outils linguistiques ne sont pas parfaitement adaptés au type de vocabulaire rencontré dans les textes spécifiques (technique, langage parlé,…). Ils génèrent une trop grande variété de termes et cela diminue l'impact statistique de chacun d'eux.

Dans cette optique, notre problématique reste moins ambitieuse puisque nous nous limitons au domaine du produit alimentaire et de ses caractéristiques sensorielles. Aussi l'utilisation de telles approches ne semble pas justifiée.

De plus, en ce qui concerne l'utilisation de tels outils informatiques pour répondre aux besoins du traitement automatique du langage naturel, nous retiendrons la réflexion de P. SAINT-DIZIER de l'Institut de Recherche en Informatique de Toulouse **[GDR-95]** :

"*Avant d'évoluer vers un nouvel outil, interrogeons-nous aussi sur son opportunité et sur le gain réel qu'il apportera à tel aspect du problème traité. Ces interrogations n'ont en général pas de réponses évidentes et souvent seule une expérimentation en profondeur pourra y répondre*."

Extraction terminologique (indexation automatique)

Le rôle principal de l'indexation est de représenter de façon condensée le document en réduisant le volume de données d'un texte.
Nous pouvons distinguer quatre types d'indexation automatique **[CHAU92]** :

- indexation par contraction et traduction de concepts qui part du langage naturel du texte du document pour aboutir à une liste de descripteurs (modèle statistique)

- indexation fondée sur le modèle probabiliste qui établit un indice de fréquence entre les termes selon la distribution des mots dans le texte

- indexation suivant le modèle linguistique (niveaux morphologique, lexical, syntaxique, sémantique et pragmatique)

- indexation par extraction qui part aussi du langage naturel du texte pour aboutir à un ensemble de mots canoniques (modèle procédural ou conceptuel)

Il est souvent question de traitement en texte intégral. Ce sont des systèmes qui prennent en compte l'intégralité des termes des textes à l'exclusion de ceux figurant dans une liste de mots vides. Pour J. CHAUMIER, il ne s'agit de *la négation même de l'indexation*.

Chaque approche est intéressante et apporte une complémentarité vis à vis des trois autres. Malheureusement, il existe très peu de système opérationnel à l'heure actuelle qui soit capable de représenter un texte à travers des règles de représentation du contenu en tenant compte du poids et du sens des mots. De plus, ces derniers sont également très complexes car ils sont destinés à traiter plusieurs domaines et plusieurs langues **[CORE94]**.

Reformatage

La bibliométrie est caractérisée comme un outil statistique de mesure de tendance de la science, des techniques et des technologies. Selon H. ROSTAING c'est "*l'application de méthodes statistiques ou mathématiques sur des ensembles de références bibliographiques*" **[ROST96]**.
Le reformatage des données est très souvent réalisée en bibliométrie pour non seulement dédoublonner l'ensemble des références bibliométriques mais également homogénéiser leur forme et leur contenu **[ROST93]**.
Concernant la forme, le reformatage supprime, crée ou réorganise les champs afin de récupérer des références issues de plusieurs sources différentes dans un format commun.
Au niveau du contenu, il standardise les champs rédigés différemment (pays, langues, auteurs, dates …) et harmonise les champs de descripteurs à l'aide de listes d'autorité.

Ce type de traitement est simple, fiable et peut être rapidement mis en place. Il est tout à fait capable de réaliser une lemmatisation automatique à partir de dictionnaires de lemmes.

Réduction de la perte d'information

Index, concordances et contexte

Grâce aux index, chacune des formes peut repérer immédiatement tous les endroits du corpus où sont situés ses occurrences. La localisation de l'ensemble des cooccurrences dans le texte d'origine est également envisageable. Les contextes immédiats des occurrences extraites peuvent être de cette façon étudiés systématiquement. Mais lorsque les fréquences ont des valeurs élevées, le travail répétitif d'examen des occurrences s'avère très contraignant.

Une autre solution est proposée en lexicométrie, il s'agit des concordances. D'une façon plus précise, une concordance représente le blocage d'une forme appelée forme pôle avec un nombre de formes défini avant et après cette forme pôle. Ces réorganisations permettent d'étudier plus facilement les rapports qui peuvent exister entre les différents contextes d'une même forme.
Cette fonctionnalité existe dans plusieurs logiciel de lexicométrie (SPAD.T, LEXICO, HYPERBASE, …) et également dans un logiciel de bibliométrie, DATAVIEW.

Ces techniques sont tout à fait intéressantes dans une phase d'évaluation, de vérification du vocabulaire ou encore comme aide à la lemmatisation par exemple pour vérifier le contexte de citation des différentes formes ambiguës. Mais elle se révèle vite fastidieuse lorsqu'elle est utilisée en routine dans le but de réduire la perte d'information.

Segments répétés

Il arrive souvent que certaines expressions ou certains groupes de mots aient un sens à part entière et que leur segmentation en mots soit une perte d'information ou une cause d'erreurs d'interprétation.
Un segment est donc une suite d'au moins deux mots répétés au moins deux fois dans l'ensemble de toutes les réponses.
Si nous rencontrons quatre fois le groupe de mots "pas très bon", c'est un segment répété de longueur trois et de fréquence quatre.

Selon L. LEBART, *leur prise en compte permet de répondre en partie aux questions concernant le choix des unités statistiques les plus pertinentes* **[LEBA88]**. Mais M. BECUE rajoute que *malheureusement, le fait qu'une même expression puisse donner lieu à de multiples segments répétés, et donc qu'un même segment répété long soit repris dans de multiples segments répétés plus courts, introduit une distorsion des résultats; de plus un certain nombre d'individus ne sont pas pris en compte car leur réponse ne contient pas de segment répété* **[BECU93]**. Pour répondre à ce problème, elle propose une autre notion : les quasi-segments.

Quasi-segments

D'une façon générale, les fréquences d'apparition des segments répétés dans un corpus sont très faibles. Ceci est dû à l'existence de modifications lexicales mineures touchant l'un des composants du segment.
Les quasi-segments rassemblent en une même unité des séquences comprenant cette unité au sein d'un intervalle défini **[BECU93]**. Par exemple, *faire sport* regroupera à la fois *faire du sport*, *faire de temps en temps du sport* **[LEBA94]**.

Cette approche augmente donc les fréquences de citation des segments, mais L. LEBART fait remarquer que *les quasi-segments sont encore plus nombreux que les segments, et leur recensement pose des problèmes de sélection et d'édition*.

Syntagmes répétés

A. PIBAROT propose une autre approche qui consiste à regrouper des groupes de mots après lemmatisation qu'il nomme syntagmes répétés **[PIBA98]**. Par exemple, *charge travail* regroupera *charge de travail*, *conséquence travail santé* et *conséquences du travail sur la santé*.
De cette façon, l'information à traiter est considérablement réduite (92%).

Cooccurrences

Nous parlerons ici de l'importance des cooccurrences dans la contextualisation. Pour le calcul statistique, nous le développerons à la section 2.1.3.
Plusieurs méthodes permettent de sélectionner le contexte de citation autour d'une forme. Nous avons déjà énoncé le cas des concordances plus haut et également de l'extraction terminologique.

La définition des unités de voisinage permet d'identifier les formes cooccurrentes dans le corpus. Il peut s'agir de la phrase entière ou d'une partie, comme dans le cas des concordances, de part et d'autre de la forme pôle.

D. LABBE montre les relations qu'il existe entre les mots par ce qu'il nomme l'univers lexical **[LABB98]**. Sans aller jusqu'à faire l'ensemble du recensement des associations entre toutes les formes d'un corpus, il divise l'ensemble des phrases en deux sous-ensembles : P1, sous-ensemble de celles qui contiennent forme1 et P0, sous-ensemble des unités desquelles forme1 est absente.
Le test de l'écart-réduit est ensuite appliqué aux autres formes du corpus sur la valeur de leurs fréquences dans chaque sous-ensemble P0 et P1 en tenant compte de leurs longueurs respectives. Dans le cas où les fréquences des formes considérées ne seraient pas trop faibles, un ensemble de formes situées dans les mêmes phrases est désigné pour chaque forme pôle donnée.
D'autres méthodes proposent des variations de celle-ci toujours dans le même optique de travail **[LEBA94]**.

L. LEBART justifie cette approche comme importante car *certaines incertitudes rencontrées lors de la lecture optique de caractères peuvent être levées (au moins en probabilité) par la considération des formes voisines déjà reconnues, si l'on connaît les probabilités d'association. La désambiguïsation lors d'une analyse morpho-syntaxique peut être réalisée dans les mêmes conditions*.
A ce sujet, il cite notamment les travaux de CHURCH et HANKS (1990). *Ces auteurs proposent d'utiliser comme mesure d'association entre deux formes x et y l'information mutuelle I(x,y), issue de la théorie de la communication de Shannon* (voir section 1.2.1) *: I(x,y) = log2 P(x,y)*
P(x)P(y) où P(x) et P(y) sont les fréquences des formes x et y dans un corpus, et P(x,y) la fréquence des occurrences voisines des deux formes, x précédant y (il n'y a donc pas symétrie vis-à-vis de x et y), le voisinage étant défini par une distance comptée en nombre de formes. Ainsi, pour les textes en anglais, ces auteurs préconisent de considérer comme voisines deux formes séparées par moins de cinq formes.

Ces dernières expériences sur les cooccurrences semblent finalement assez adaptées à notre recherche de réduction de la perte d'information. C'est dans cet esprit que nous allons maintenant étudier les différentes possibilités de traitements statistiques et de représentation graphique.

2.1.3. Sur le traitement statistique et la représentation graphique

Nous avons divisé le traitement statistique des données textuelles en trois catégories différentes et complémentaires :

- l'analyse lexicale
- l'analyse multidimentionnelle
- l'analyse des cooccurrences

Elles sont bien sûr choisies suivant les objectifs de l'étude. Elles permettent donc d'aborder différemment des données textuelles identiques et proposent plusieurs mode de représentation plus ou moins simples à interpréter. Nous aborderons plus en détail les deux premières catégories aux sections 2.2.1 et 2.2.2 à travers leur utilisation sur les commentaires libres de consommateurs. Par contre, la troisième correspond à une nouvelle approche intéressante que nous souhaitons développer.

Analyse lexicale

Il s'agit d'un simple comptage de mots, des couples de mots ou des segments répétés à partir du lexique du corpus.
Elle permet de dégager les mots, couples de mots, segments répétés les plus significatifs. Par exemple, on peut mettre en évidence que la note caramel est plus citée pour la bière *1664 brune* que pour les autres bières du marché.

Analyse multidimentionnelle

Il s'agit d'analyse factorielle des correspondances et de classifications hiérarchiques (ascendante ou descendante).
Elle permet de partitionner les données en classe thématiques et de les positionner les unes par rapport aux autres. Par exemple, on pourra mettre en évidence un jugement différent entre les hommes et les femmes en interprétant le plan factoriel et la CAH (voir Figure 15, page 79).

Analyse des cooccurrences

Il s'agit de graphes de connexion représentant les co-apparitions des mots dans les unités de décompte choisies.
Elle permet d'identifier les principales associations des mots représentant les thèmes globaux présents dans le corpus de données **[COUR94]**. Par exemple, on peut mettre en évidence des liens entre le désaltérant et l'amertume d'une bière. Ce lien permettra de construire l'hypothèse que l'amertume peut expliquer le caractère désaltérant.
Ce type d'analyse est très utilisé en bibliométrie pour cartographier les travaux scientifiques d'un auteur ou groupe d'auteurs à partir du titre ou des mots clés contenus dans les publications **[ROST96]**.

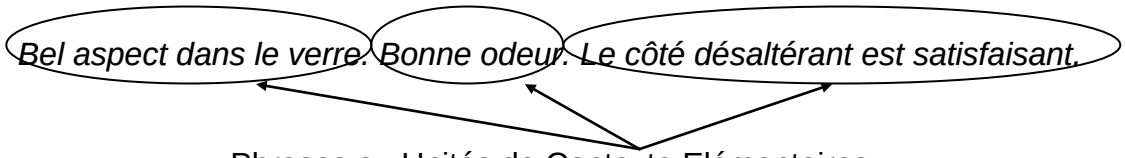
Les mots sont à l'origine liés par des règles grammaticales, sémantiques et pragmatiques. Ils sont ensuite isolés et comptabilisés dans une finalité de traitement statistique.
Le graphe de connexion permet de retrouver les liaisons les plus fortes et les plus fréquentes existantes entre les mots. Nous retrouvons cette idée dans **[DUMA94]**, *le contenu informationnel d'une paire de mots est plus important que celui de deux mots pris séparément*.

Ce mode de calcul nous semble intéressant pour traiter les commentaires libres. Aussi avons nous choisi d'étudier de manière plus approfondie les éléments sur lesquels s'effectueront les calculs ainsi que les modes de calculs qui nous permettront de visualiser l'information de façon pertinente.

Choix des associations

Avant de pratiquer des comptages sur les textes, il est nécessaire de procéder à une segmentation de la chaîne textuelle – phrases - en unités distinctes - données textuelles exploitables - (voir sections 1.2.1 et 2.1.2).
Nous devons déterminer l'ensemble des mots sur lequel nous choisissons d'effectuer les calculs.
L. LEBART et A. SALEM, choisissent de travailler sur les formes graphiques définies comme des suites de caractères comprises entre deux caractères délimiteurs **[LEBA88]**.
M. REINERT considère qu'une réponse ou un commentaire est une Unité de Contexte Initiale (voir section 2.2.2) **[REIN86]**. Il est composé de plusieurs Unités de Contexte Elémentaires dont la taille est variable.
Pour G. TEIL, il s'agit d'unités de sens **[TEI92b]**.

Si nous voulons identifier les principaux thèmes qui se dégagent dans le discours des consommateurs pour décrire un produit, nous choisirons de calculer les cooccurrences à partir des unités élémentaires. Elles correspondent à la chaîne de caractère de taille variable qui est comprise entre deux points.
EXEMPLE :

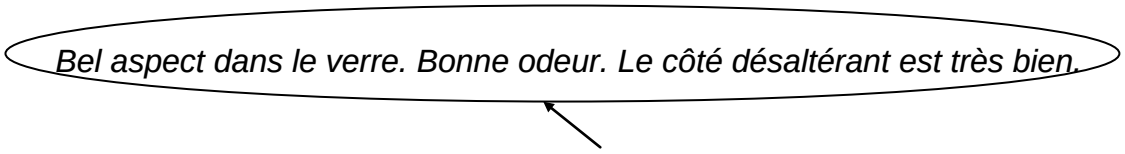


Phrases ou Unités de Contexte Elémentaires

Avec ce découpage, nous obtiendrons les paires suivantes :

beau ----- aspect
bon ----- odeur
désaltérant ----- satisfaisant

Par contre, si nous voulons identifier les descripteurs qui sont associés dans le discours des consommateurs (par exemple, *l'amertume est fortement associée au désaltérant et au goût*), nous choisirons de calculer les cooccurrences dans le commentaire entier.
EXEMPLE :

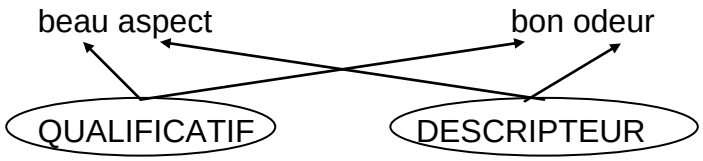


Commentaire ou Unité de Contexte Initiale

Sans découpage mais avec la reconnaissance des descripteurs *aspects*, *odeur* et *désaltérant*, nous obtiendrons les paires suivantes :

aspect ----- odeur
aspect ----- désaltérant
odeur ----- désaltérant

Le deuxième type d'association, sous-entend qu'il sera nécessaire de distinguer les qualificatifs des descripteurs, des descripteurs eux-mêmes.
EXEMPLE :



Un codage spécifique doit mettre en évidence ces deux types de termes. Le traitement statistique portera alors uniquement sur les descripteurs à l'intérieur d'un même commentaire.

Quelques modes de calcul des paires de mots

L'apparition simultanée de deux éléments (des formes descriptives en ce qui nous concerne) peut être calculée de plusieurs façons différentes. Ces différents calculs seront classés en fonction de leurs propriétés mathématiques :

La fréquence

La fréquence d'apparition d'une paire permet de mettre très rapidement en évidence les liens qui existent entre les différents mots mais elle présente l'inconvénient d'égaliser le poids de toutes les entités. En effet, elle ne tient pas compte des fréquences relatives de deux mots **[DUMA94]**.

La fréquence tient compte du poids du lien mais pas de son intensité.

Les indices d'association

Ils mesurent la ressemblance entre deux mots suivant un calcul basé sur les données binaires de présence-absence. Nous utiliserons dans les lignes qui suivent, les variables exprimées dans le Tableau 13 et la Figure 7 suivants :

Tableau 13 : Présence/Absence des mots X et Y

		Mot Y	
		Présence	**Absence**
Mot X	**Présence**	Na	Nb
	Absence	Nc	Nd

Avec :

Na = Nombre d'apparition de la paire (double présence ou cooccurrence)
Nb et Nc = Nombre d'apparition du mot seul
Nd = Nombre de non apparition de la paire (double absence)

M = Na + Nb + Nc + Nd = Nombre total de commentaires

Na + Nb = Nombre d'apparition du mot X (occurrence)
Na + Nc = Nombre d'apparition du mot Y (occurrence)

Na et Nd caractérisent la similitude entre les deux mots,
Nb et Nc caractérisent la dissimilitude entre les deux mots.

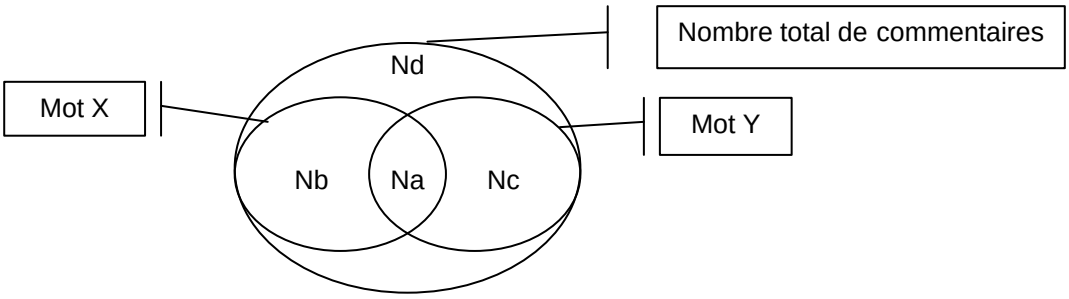


Figure 7 : Présence/Absence des mots X et Y

A partir de ces éléments, nous pouvons calculer n'importe quel coefficient ou indice d'association entre deux mots X et Y contenus dans le corpus de commentaires.

Selon F. MARCOTORCHINO, nous pouvons classer ces indices en quatre grands types **[MARC81]** :

① Les indices qui favorisent Na et Nd et qui défavorisent Nb et Nc

EXEMPLE : L'indice de Sokal et Michener

Formule 1 : (Na + Nd)/(Na + Nb + Nc)

② Les indices qui favorisent Na et qui défavorisent Nb, Nc et Nd

EXEMPLE : L'indice de Russel et Rao

Formule 2 : Na/(Na + Nb + Nc + Nd)

③ Les indices qui favorisent Na et qui défavorisent Nb et Nc sans favoriser ni défavoriser Nd
EXEMPLE : L'indice de Marcotorchino-Michaud

Formule 3: (Na + ½ Nd)/(Na + Nb + Nc + Nd)

④ Les indices qui favorisent Na et qui défavorisent Nb et Nc sans considérer Nd

EXEMPLE : L'ındice de Jaccard

Formule 4 : Na/(Na+ Nb + Nc)

① et ③ considèrent que la double absence est un facteur de ressemblance entre les mots.
② représente une fréquence relative à la taille du corpus.
④ favorise les paires fortement liées même si leur fréquence est faible.

Les indices de similitude

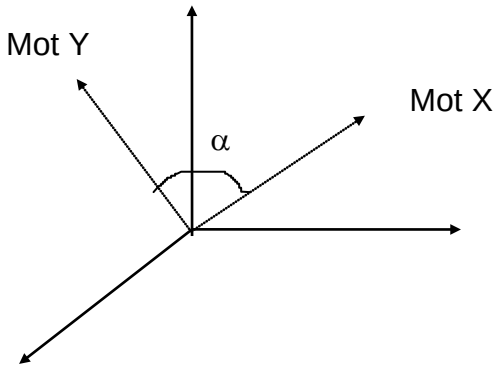
Nous prenons en considération l'information similaire c'est-à-dire l'ensemble des cas où les mots sont co-présents ou co-absents.
Deux cas sont envisagés :

× La prise en compte de la double absence

EXEMPLE : Le coefficient de corrélation

Formule 5 : $((Na * Nd)-(Nb * Nc)) / \sqrt{((Na + Nb)*(Na + Nc)*(Nb + Nd)*(Nc + Nd))}$

Il varie de - 1 à 1. C'est le cosinus de l'angle formé par les deux vecteurs :



Si cosα = 1, l'angle est nul et les deux vecteurs sont colinéaires donc les mots sont fortement liés.

Si cosα = 0, l'angle est droit et les deux vecteurs sont indépendants (au sens statistique) donc les mots ne sont pas liés.

Si cosα = -1, l'angle est de 180° et les deux vecteurs sont opposés donc les mots sont dans des espaces lexicaux totalement opposés donc ils ne devraient pas entretenir de relation.

× Pas de prise en compte de la double absence

EXEMPLES : ① L'indice de Jaccard (voir Formule 1)

Il varie de 0 à 1. Il peut être assimilé à un pourcentage. Il a donc directement un sens pour l'interprétation. C'est le coefficient le plus utilisé en calcul de cooccurrence en bibliométrie. Il est très peu différent du coefficient d'équivalence (voir plus bas pour la définition). En effet, ils varient tous les deux de la même manière.

② L'indice d'inclusion

Formule 6: $Na / \min \{(Na + Nb),(Na + Nc)\}$

Il varie de 0 à 1. C'est un coefficient de similitude un peu particulier. En effet, il n'est pas symétrique suivant les apparitions d'occurrence puisqu'il ne prend en compte que le nombre d'occurrence le plus petit.

③ Le coefficient d'équivalence

Formule 7: $Na^2 / \{(Na + Nb) * (Na + Nc)\}$

Il varie de 0 à 1. Il mesure l'exclusivité de l'association de deux mots. C'est un coefficient local et homogène. C'est l'indice qui est employé dans les algorithmes des logiciels LEXIMAPPE™ et CANDIDE™ (voir section 2.2.3).

Les indices de dissimilitude

Nous prenons en considération l'ensemble des cas où les mots sont présents de façon isolée et non par paire. Par opposition aux indices de similitude, les valeurs des indices de dissimilitude augmentent lorsque la dissemblance entre les mots est plus grande. Pratiquement, cela signifie que lorsque l'indice est égal à 1, les deux mots sont très distants dans le cas de la dissimilitude alors qu'ils sont très proches dans le cas de la similitude.

Ici aussi, deux cas sont envisagés :

- Pas de prise en compte de la double absence

EXEMPLE : Le coefficient de Bray & Curtis

Formule 8: (Nb + Nc) / (2Na + Nb + Nc)

Il varie de 0 à 1.

- La prise en compte de la double absence

EXEMPLE : La distance euclidienne binaire

Formule 9: (Nb + Nc) / (Na + Nb + Nc + Nd)

Elle varie de 0 à 1.

Comparaison d'indices

La littérature propose, nous venons de le voir une liste de coefficients pour le calcul des cooccurrences. Ils ont bien entendu leurs spécificités puisqu'ils ont chacun été établis dans le but de mettre en évidence un type d'information précis.
Afin de choisir la méthode de calcul de cooccurrence la plus adaptée à nos besoins, nous avons mis à l'épreuve quatre d'entre eux.
Nous avons donc choisi de comparer l'indice de Jaccard, l'indice d'inclusion (ou indice d'inclusion réciproque selon B. MICHELET, la corrélation et l'indice d'équivalence en fonction de leurs effets **[MICH88]**. Il s'agit là de quatre indices de similitude puisque nous nous intéressons plus précisément à la dépendance des mots. De plus, le fait que les indices de dissimilitude aient pour effet d'isoler les mots de forte fréquence et de créer des liens multiples par effet de chaînage, nous conforte dans cette idée **[DUMA94]**.
Ces quatre indices sont locaux puisqu'ils ne font pas intervenir le nombre total de commentaires[16]. Nous pourrons donc selon B. MICHELET établir une comparaison à travers leurs effets **[MICH88]**.

Plusieurs cas sont étudiés pour évaluer les variations des différents indices pour un même corpus donné de 1000 commentaires. Nous cherchons à observer pour chaque indice, ses variations en fonction des valeurs des cooccurrences (Na), des occurrences (Nc et Nb) et des absences (Nd). Dans tous les cas, nous prendrons les valeurs extrêmes (minimum et maximum).

[16] Ce n'est pas le cas de la distance euclidienne

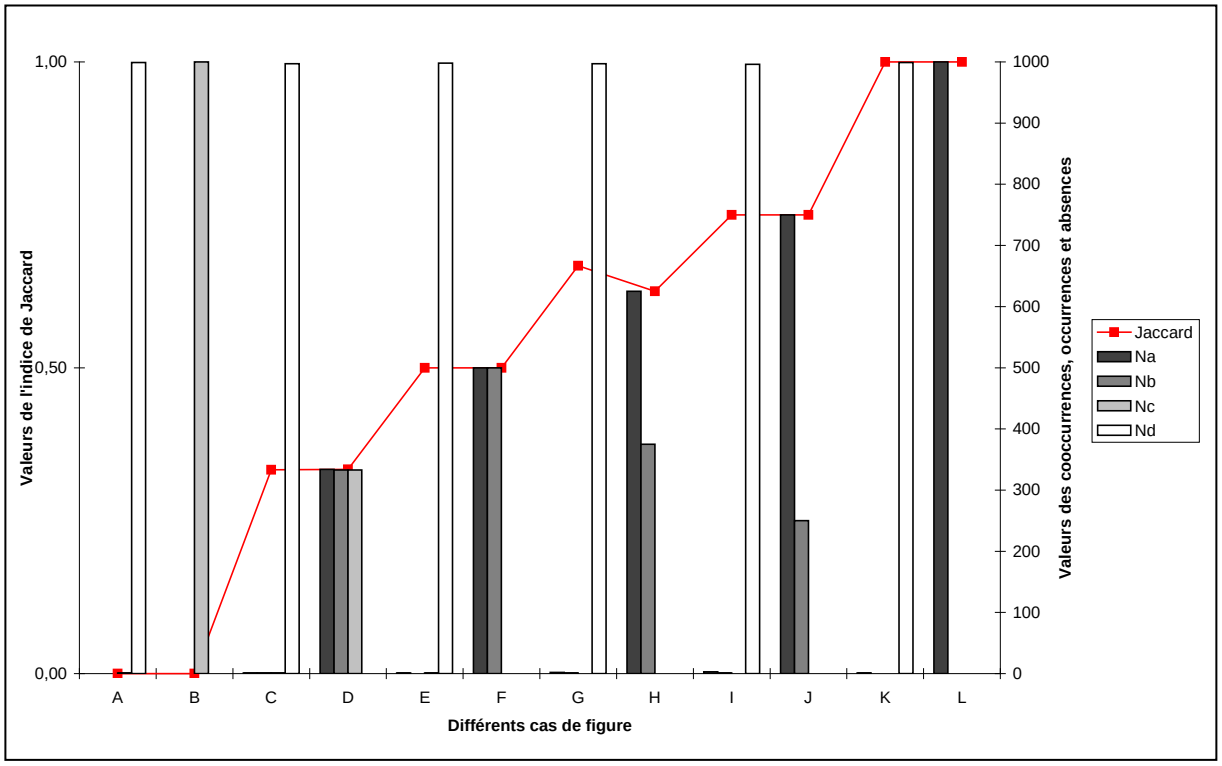


Figure 8 : Variation de l'indice de Jaccard en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Tableau 14 : Valeurs de l'indice de Jaccard en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Cas	A	B	C	D	E	F	G	H	I	J	K	L
Na	0	0	1	334	1	500	2	625	3	750	1	1000
Nb	0	0	1	333	0	500	1	375	1	250	0	0
Nc	1	1000	1	333	1	0	0	0	0	0	0	0
Nd	999	0	997	0	998	0	997	0	996	0	999	0
Jaccard	0,00	0,00	0,33	0,33	0,50	0,50	0,67	0,63	0,75	0,75	1,00	1,00

L'observation la plus frappante sur la Figure 8 est le parallélisme entre l'augmentation de l'indice de Jaccard et l'augmentation des valeurs des cooccurrences.

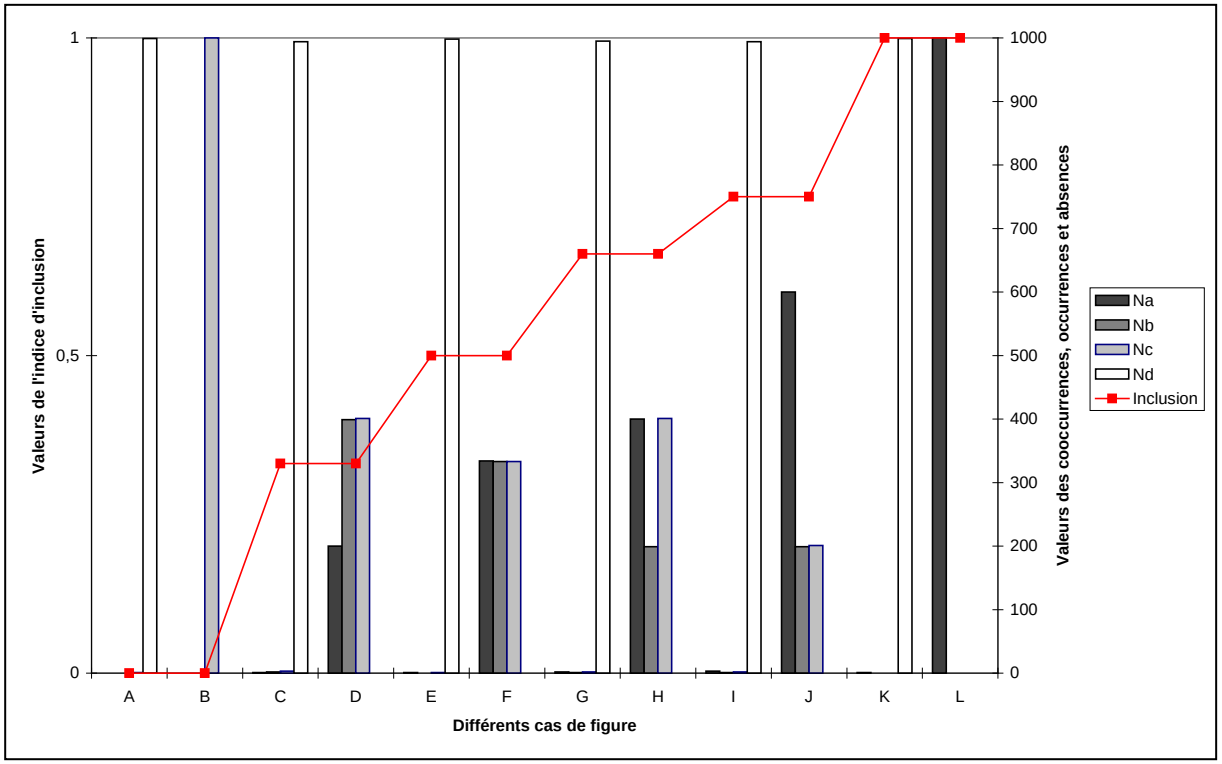


Figure 9 : Variation de l'indice d'inclusion en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Tableau 15 : Valeurs de l'indice d'inclusion en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Cas	A	B	C	D	E	F	G	H	I	J	K	L
Na	0	0	1	200	1	334	2	400	3	600	1	1000
Nb	0	0	2	399	0	333	1	199	1	199	0	0
Nc	1	1000	3	401	1	333	2	401	2	201	0	0
Nd	999	0	994	0	998	0	995	0	994	0	999	0
Inclusion	0	0	0,33	0,33	0,5	0,5	0,66	0,66	0,75	0,75	1	1

Nous observons également un parallélisme entre les cooccurrences et l'indice d'inclusion sur la Figure 9 mais de façon moins régulière.
Les variations de l'indice de Jaccard et de l'inclusion sont très ressemblantes.

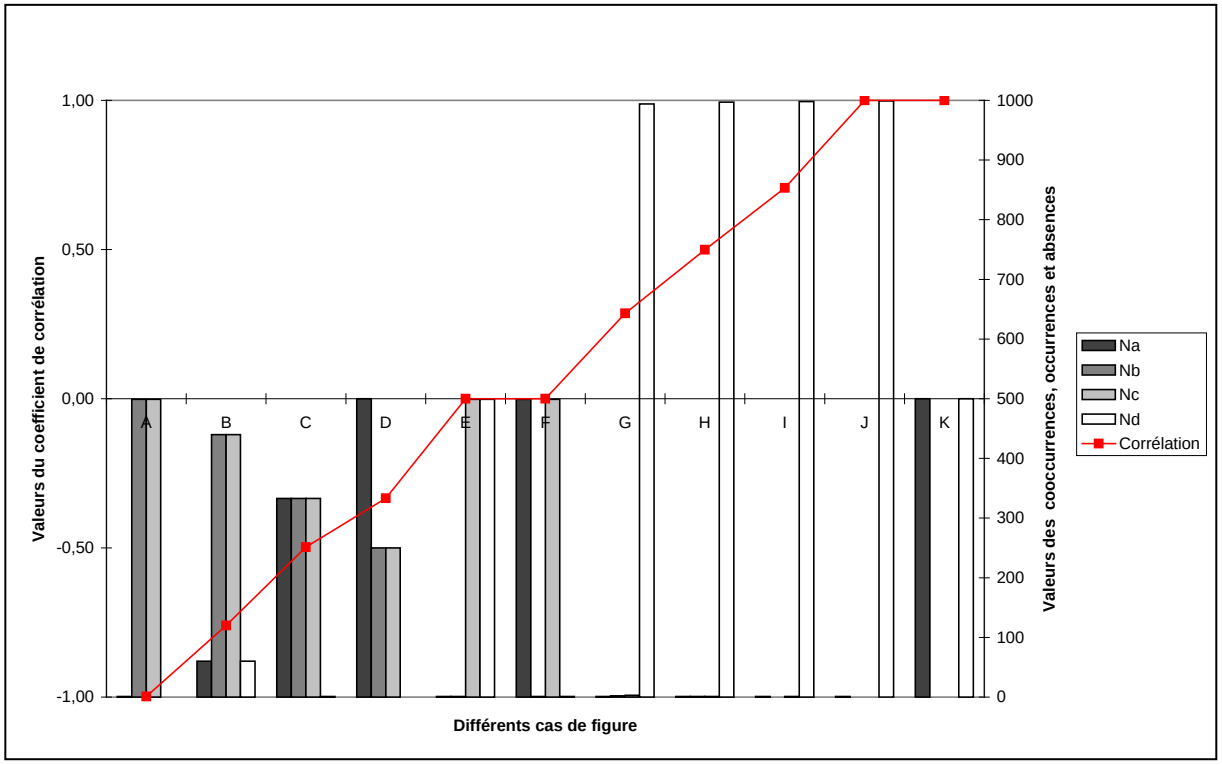


Figure 10: Variation du coefficient de corrélation en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Tableau 16 : Valeurs du coefficient de corrélation en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Cas	A	B	C	D	E	F	G	H	I	J	K
Na	1	60	334	500	1	499	1	1	1	1	500
Nb	499	440	333	250	1	1	2	1	0	0	0
Nc	499	440	333	250	499	499	3	1	1	0	0
Nd	0	60	0	0	499	1	994	997	998	999	500
Corrélation	-1,00	-0,76	-0,50	-0,33	0,00	0,00	0,29	0,50	0,71	1,00	1,00

Le coefficient de corrélation varie surtout en fonction des valeurs des occurrences Nc et Nb.

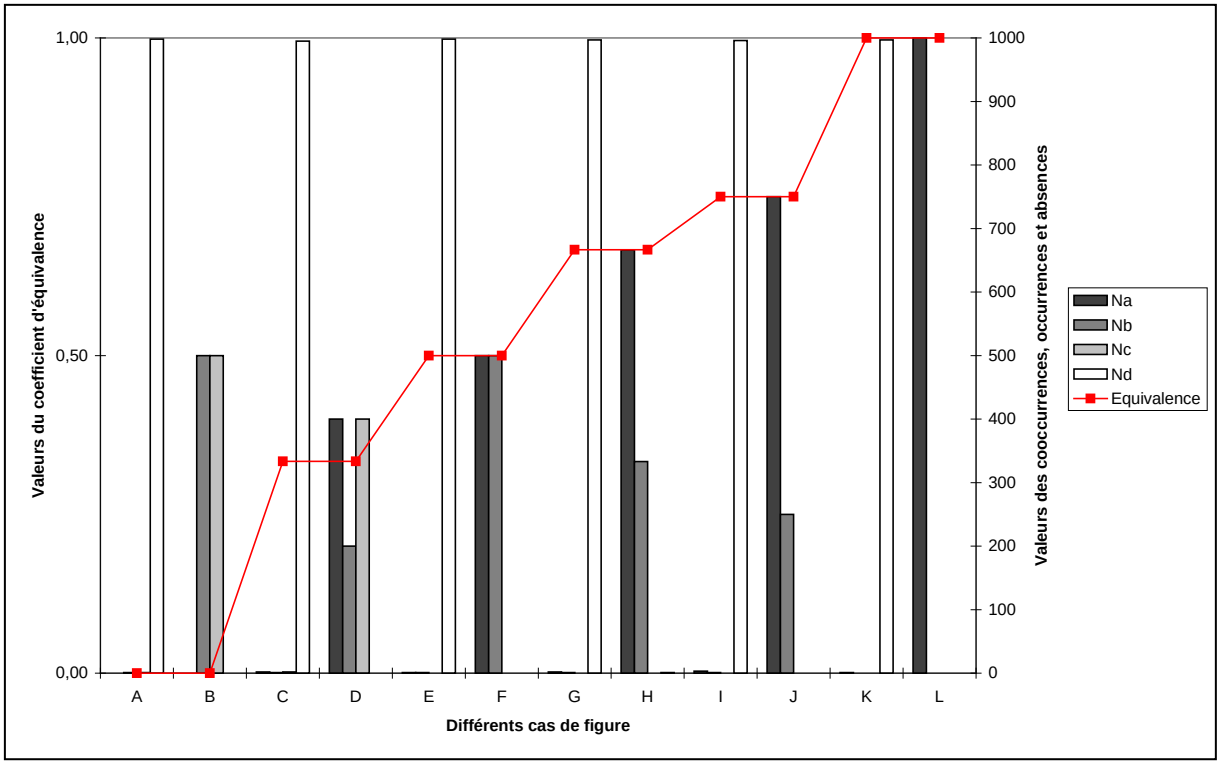


Figure 11 : Variation du coefficient d'équivalence en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Tableau 17: Valeurs du coefficient d'équivalence en fonction de différentes valeurs de cooccurrence, d'occurrence et d'absence

Cas	A	B	C	D	E	F	G	H	I	J	K	L
Na	0	0	2	400	1	500	2	666	3	750	1	1000
Nb	1	500	1	200	1	500	1	333	1	250	0	0
Nc	1	500	2	400	0	0	0	0	0	0	0	0
Nd	998	0	995	0	998	0	997	1	996	0	997	0
Equivalence	0,00	0,00	0,33	0,33	0,50	0,50	0,67	0,67	0,75	0,75	1,00	1,00

La variation du coefficient d'équivalence est pratiquement identique à celle de l'indice de Jaccard, du moins en ce qui concerne les valeurs des cooccurrences.
Nous remarquons une différence avec la variation de l'indice de Jaccard au niveau des valeurs des occurrences de mots isolés.
Par exemple une valeur d'indice égale à 0,33 sera due à la présence d'autant de co-apparitions que d'apparition des deux mots sous forme isolée pour l'indice de Jaccard. Alors que pour l'équivalence elle sera due à la présence d'autant de co-apparitions que d'apparition d'un des deux mots sous forme isolée et à la moitié de l'autre mot sous forme isolée.

Jaccard est donc un indice qui favorise les paires fortement liées même si leur fréquence d'apparition est faible.
Il s'intéresse exclusivement à la co-présence des mots à l'intérieur d'un commentaire pour calculer l'association.
B. MICHELET montre sur les triangles de représentation, que Jaccard met en évidence les liaisons entre des groupes de mots de taille équivalente **[MICH88]**. Il le qualifie de bon indice pour représenter les groupes de mots co-cités avec des fréquences proches (faibles ou fortes).
En définitive, c'est un excellent indice d'association tant qu'il n'est pas utilisé comme une distance. B. MICHELET montre que la nature de son dénominateur provoque certaines inégalités **[MICH88]**.

L'inclusion favorise les paires dont les mots n'apparaissent qu'associés et ceci est valable même si leur fréquence d'apparition est faible.
De la même façon que pour l'indice de Jaccard, B. MICHELET montre que l'indice d'inclusion est asymétrique **[MICH88]**. Les liaisons entre les mots seront alors orientées. D'autre part, il assimile l'inclusion à la fréquence relative qui varie en fonction de la fréquence de mots cooccurrents.

La corrélation considère l'absence simultanée et les apparitions de mots isolés comme des facteurs importants de la ressemblance entre deux mots.
Elle met donc en évidence une autre partie de la représentation. Cette partie est plutôt considérée comme supplémentaire car la variation de la corrélation n'est pas comparable aux variations des indices précédents puisqu'elle peut prendre des valeurs négatives.
Avec la corrélation, nous mettrons en évidence les paires dont les mots constituants ne sont présents que très souvent ensemble. Lorsqu'un des deux mots est fortement présent de façon isolée, l'indice passe en négatif. Ceci signifie que les deux mots appartiennent à des ensembles différents.
Le coefficient de corrélation est le seul qui marque aussi bien la différence entre les cooccurrences et les occurrences.

Les effets de l'indice d'équivalence sont comparables à l'indice de Jaccard mais uniquement sur les fréquences fortes. C'est l'indice de référence pour les calculs de cooccurrences car il est homogène[17] et peu être assimilé à une distance[18]. Cette dernière propriété n'est pas justifiée dans notre approche car d'une manière générale, la notion de distance n'a pas de sens pour les analyses des associations. Elle permet seulement de fournir une aide à la lecture des résultats pour les esprits sensibilisés aux règles statistiques.

Enfin, l'équivalence est non seulement étroitement liée à l'indice de Jaccard pour son profil de variation identique (voir les Figure 8 et Figure 11) mais aussi à l'indice d'inclusion pour sa formule de calcul.

[17] Lorsque nous ajoutons dans notre corpus un ensemble de commentaires voisin, le calcul du réseau ne sera pas modifié
[18] Pratique si nous souhaitons une disposition spatiale

En effet, si nous partons de la formule de l'indice d'inclusion :

$$\frac{Na}{\min\{(Na + Nb),(Na + Nc)\}}$$

En choisissant de partir de l'hypothèse que Nb est toujours inférieur à Nc, alors :

$$\frac{Na}{(Na + Nb)}$$

Si nous élevons ce coefficient au carré, nous aurons :

$$\frac{Na^2}{\{(Na + Nb) * (Na + Nb)\}}$$

Nous observons donc bien une similitude dans la formule à la valeur d'un des mots qui apparaît seul près (voir Formule 7). Cette remarque est confirmée par le fait que l'indice d'équivalence soit le produit de l'indice d'inclusion par l'indice d'inclusion réciproque **[MICH88]**.

Lorsque nous observons la variation des trois coefficients (Jaccard, Inclusion et Equivalence) en fonction des valeurs des cooccurrences sur la Figure 12, nous voyons l'indice d'inclusion évolue avec les valeurs les plus faibles :

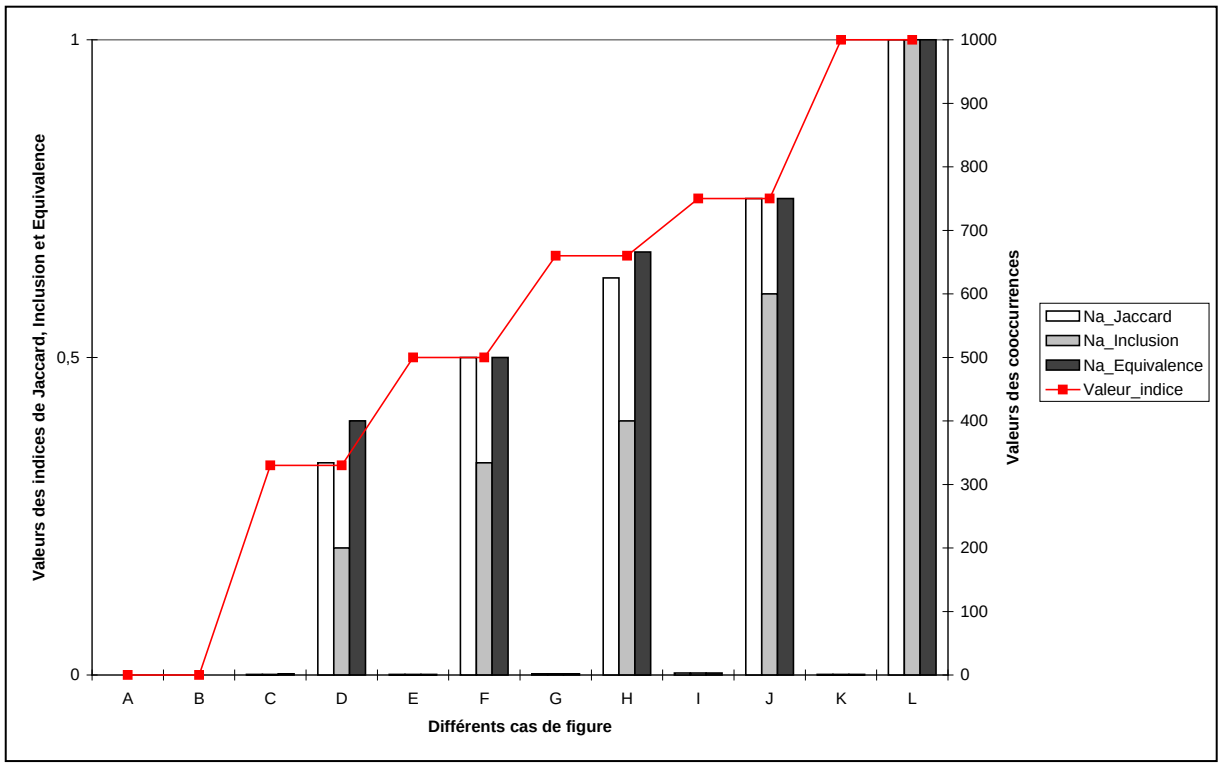


Figure 12 : Variation des coocurrences pour différentes valeurs des indices de Jaccard, Inclusion et Equivalence

En conclusion, cette comparaison nous a permis de mettre en évidence des effets intéressants sur les variations de certains indices de similitude. D'autres études (**[MICH88], [CAIL76], [BEN73a], [BEN73b]**) nous ont apporté des éléments supplémentaires pour mieux les connaître.
Nous avons trouvé les indices d'inclusion, de Jaccard et de corrélation intéressants pour leurs effets complémentaires et parce qu'ils sont en adéquation avec la particularité de nos données.

2.2. Dans l'expérience DANONE

Le projet d'étude du vocabulaire et des expressions employées par le consommateur a commencé il y a 6 ans au TEPRAL. Quatre types de logiciels ont été utilisés pour mener à bien le traitement de ces données :

- SPADT ou Système Portable pour l'Analyse des Données Textuelles du CISIA
- ALCESTE ou Analyse des Lexèmes Cooccurrents dans les Enoncés Simples d'un Texte de la société IMAGE
- CANDIDE™ du Centre de Sociologie de l'innovation de l'Ecole des Mines de Paris.
- LEXICO[19] du laboratoire Lexicométrie & textes politiques de l'E.N.S. de Fontenay-Saint-Cloud.

Le traitement des données textuelles à l'aide de ces logiciels a été réalisé avant le début de cette thèse. Nous présentons donc ici quelques synthèses d'études auxquelles nous n'avons pas participé[20].

2.2.1. L'analyse des données textuelles avec SPADT

Présentation de l'outil

Créé en 1989 au C.I.S.I.A par L. LEBART, A. MORINEAU et M. BECUE, SPAD.T est un logiciel autonome d'analyse des données textuelles **[LEB93a]**.
Il est composé de 23 procédures dont chacune représente soit une étape logique de l'analyse textuelle soit une étape de gestion de données (lecture, archivage ...).
Ce logiciel lit en entrée un fichier de données textuelles où sont stockées les réponses d'un certain nombre d'individus à des questions ouvertes, et éventuellement un fichier contenant des variables nominales décrivant ces mêmes individus.
Chaque procédure est définie par des paramètres que l'utilisateur doit fixer suivant la nature de ses données et l'étude qu'il veut effectuer.

[19] Réalisé par le centre de recherche Jean Thèves de la Branche BISCUITS du Groupe DANONE (voir section 2.4)
[20] Cependant, les approches ont été exposées dans une communication **[ZIEG96]**

C'est le premier outil qui a permis de travailler sur les données textuelles au TEPRAL. Il a été introduit par "affiliation" puisque SPAD.N a d'abord été introduit pour traiter les données numériques des tests de dégustation. SPAD.T a en quelque sorte influencé la mise en place de questions ouvertes dans les tests consommateurs. La première application a été réalisée en collaboration avec D. GANGE du CNRS de Strasbourg. Avec l'aide des études du CREDOC sur le comportement alimentaire des français (**[AUCO91]**, **[BEAU95]**, **[BEA93a]**, **[BEAU94]**, **[LAHL92]**, **[LAHL93]**, **[LION91]**, **[YVON90]**), le TEPRAL s'est réellement investi dans le domaine de la lexicométrie.

Numérisation, comptage et tableaux lexicaux

La première étape à mettre en œuvre lors d'une étude textuelle est de segmenter l'ensemble du corpus, c'est-à-dire de le décomposer en formes graphiques.
Il est alors possible d'établir un dictionnaire de tous les mots utilisés et d'opérer des comptages. Cette liste est alors rangée par ordre alphabétique ou par ordre de fréquences et chaque mot se voit attribuer un numéro (ou un rang) d'ordre. Par exemple, par ordre alphabétique, au mot "a" sera attribué le numéro 1.

A partir de là, un tableau E, appelé Tableau Lexical Entier est construit :

Tableau 18 : Exemple de tableau lexical entier

	1er mot du dictionnaire	2ème mot du dictionnaire	3ème mot du dictionnaire	…
Réponse de l'individu 1	0	2	0	…
Réponse de l'individu 2	1	0	3…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
Réponse de l'individu n	…	…	…	…

Le chiffre de l'intersection entre la ligne i et de la colonne j représente donc le nombre de fois où le mot j aura été utilisé dans la réponse du consommateur i.

Il semble évident que ce tableau peut prendre des dimensions très grandes et qu'il sera probablement composé d'un fort pourcentage de 0 (voir problème des matrices creuses à la section 2.1.2). Ceci se vérifie d'autant mieux si les réponses aux questions sont courtes (de l'ordre d'une phrase ou deux).

Plutôt que de considérer la réponse d'un consommateur comme étant individuelle, il peut être fort intéressant de regrouper ces réponses suivant une variable nominale ou un croisement de variables (selon des classes d'âge ou des CSP[21] par exemple). Cela implique l'existence dans notre questionnaire d'une partie de description de certaines caractéristiques des individus parallèlement aux questions ouvertes.

[21] Catégories Socio-Professionnelles : regroupement d'enquêtés suivant leur insertion socio-économique

Si ces variables existent, nous pouvons donc obtenir un autre tableau beaucoup plus concis que nous appelons Tableau Lexical Agrégé (A) :

Tableau 19 : Exemple de tableau lexical agrégé

	Variable ou croisement choisi			
	Modalité 1	**Modalité 2**	**Modalité 3**	…
mot 1 (numéro d'ordre 1)	2	0	5	…
mot 2 (numéro d'ordre 2)	7	3	3	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
mot m (numéro d'ordre m)	0	10	6	…

L'intersection de la ligne i et de la colonne j représente le nombre de fois où le mot i a été cité par tous les individus appartenant à la modalité j.

L'intérêt statistique de ces deux tableaux est de pouvoir leur appliquer une Analyse Factorielle des Correspondances (AFC) car ce sont des tableaux de contingence. Nous pouvons donc calculer, comparer et représenter les distances entre profils lexicaux pour les mots d'une part, et pour les réponses individuelles ou les modalités d'autre part.

Analyse lexicale

Afin de diminuer la dispersion du vocabulaire, le logiciel SPAD T propose de faire préalablement une sélection des mots (ou des segments répétés, voir section 2.1.2) selon leur fréquence et dans certains cas une suppression des mots outils (voir section 2.1.2).

L'ensemble du vocabulaire représentatif en terme statistique peut être étudié par des calculs probabilistes sur les fréquences des mots ou des réponses : ce sont les mots caractéristiques et les réponses modales**.**

A l'aide de tests de comparaison entre la fréquence globale d'un mot (dans l'ensemble des réponses) et la fréquence de ce même mot dans une partie des réponses, nous pouvons établir, pour chaque modalité, une liste des mots significativement les plus employés et les moins employés par ce groupe d'individus.

Suivant le même principe, il est possible également de calculer les phrases les plus caractéristiques pour chaque groupe d'individus, appelées réponses modales. Ces phrases ne sont en aucun cas reconstituées mais il s'agit bien des réponses authentiques, prises dans leur intégralité.

Nous les calculons de deux façons : selon la fréquence des mots, qui favorise les réponses courtes ou selon la distance du Chi-2, qui favorise plutôt les réponses longues.

La première méthode s'effectue en classant d'abord les mots d'un regroupement de réponses par ordre de signification en leur attribuant, un rang. Nous calculerons ensuite pour chaque réponse le rang moyen des mots qui la composent. Plus ce rang est petit, plus la réponse est caractéristique.
La seconde méthode compare le profil d'une réponse avec le profil moyen de la classe à laquelle elle appartient : nous utilisons pour cela la distance du Chi-2. Plus la distance est faible, plus le profil lexical de la réponse se rapproche du profil lexical du regroupement de réponses qui lui correspond, et donc plus elle caractérise cette classe.

Analyse multidimentionnelle

SPAD.T propose d'effectuer une AFC sur un des tableaux lexicaux vus plus haut (Tableau 18 et Tableau 19). Nous verrons dans les paragraphes suivants que le choix du tableau entraîne deux analyses très différentes l'une de l'autre.

L'analyse peut également se poursuivre par des Classifications Ascendantes Hiérarchiques (CAH) pour observer l'agrégation des mots ou des modalités. Le plus souvent, la classification des modalités est la plus intéressante.
Par exemple, si ce sont différentes marques de produits alimentaires, la façon dont elles se regroupent entre elles peut apporter de précieuses informations.

Application

Nous voudrions tester l'influence et l'impact de nombreux facteurs (sexe, âge, produit, combinaisons de variables, …) sur les commentaires libres de consommateurs puisque le logiciel a été conçu dans cet esprit. Les méthodes que nous allons mettre en œuvre sont alors conduites dans le but de mettre en évidence des différences de contenu dans les réponses libres en fonction des variables nominales choisies.
Cependant, nous avons voulu voir également comment SPAD.T se comportait si nous ne faisions pas intervenir les variables nominales sur le tableau lexical entier. Nous avons donc tenté deux approches différentes :

- l'analyse directe à partir du tableau lexical entier E (voir Tableau 18). Les réponses n'ont pas de regroupement a priori.
- l'analyse par regroupement sur le tableau lexical agrégé A (voir Tableau 19). Les réponses sont regroupées suivant des modalités.

<u>Présentation du corpus</u>

80 consommateurs ont dégusté en aveugle 8 bières de 6 marques différentes. Le corpus comprend donc 640 réponses. Nous distinguons deux types d'informations recueillies :

- une information provenant de questions fermées : le sexe, l'âge, la consommation de bière, le lieu de consommation, une note hédonique et l'intention d'achat.
- une information provenant de questions ouvertes : les commentaires libres sur l'odeur, le goût et l'impression globale.

Analyse directe

Une sélection préalable a été faite sur le vocabulaire. Nous avons retenu les 100 formes les plus importantes et fréquentes.
Le tableau lexical entier est composé en lignes par les réponses des consommateurs pour une bière donnée et en colonne par les formes sélectionnées. Il représente une matrice de grande taille comportant un grand nombre de zéros (environ 95%). Aussi, l'AFC comportera 99 axes.

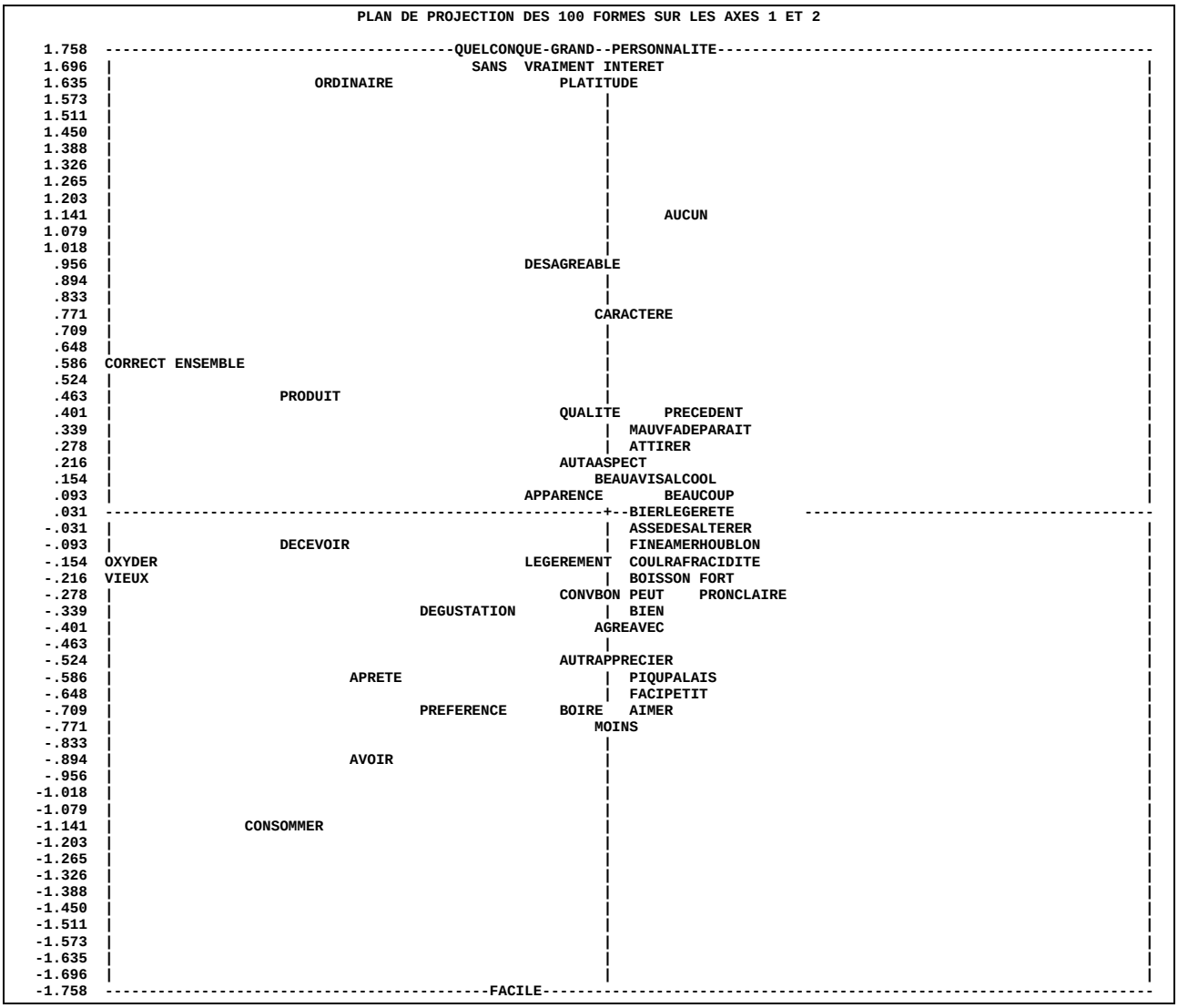


Figure 13 : Analyse factorielle des correspondances des mots sur le tableau lexical entier

La Figure 13 montre que le premier facteur (2,72% de l'inertie totale) semble assez trivial. Seuls quelques formes le composent : nous notons les formes *oxyder* (45,5% de l'inertie de l'axe) et *vieux* (28%) qui ont dû être utilisés ensemble.
Le second facteur (2,41%) paraît plus intéressant. Nous pouvons pratiquement reconstituer les réponses ou partie de réponses *facile à boire*, *agréable à boire* dans la partie inférieure et *sans grand intérêt*, *aucun caractère* accompagnées des formes *ordinaire*, *quelconque*, *désagréable* dans la partie supérieure.

Nous avons également effectué une sélection des 115 segments répétés les plus fréquents avant de lancer une analyse directe sur les segments répétés.

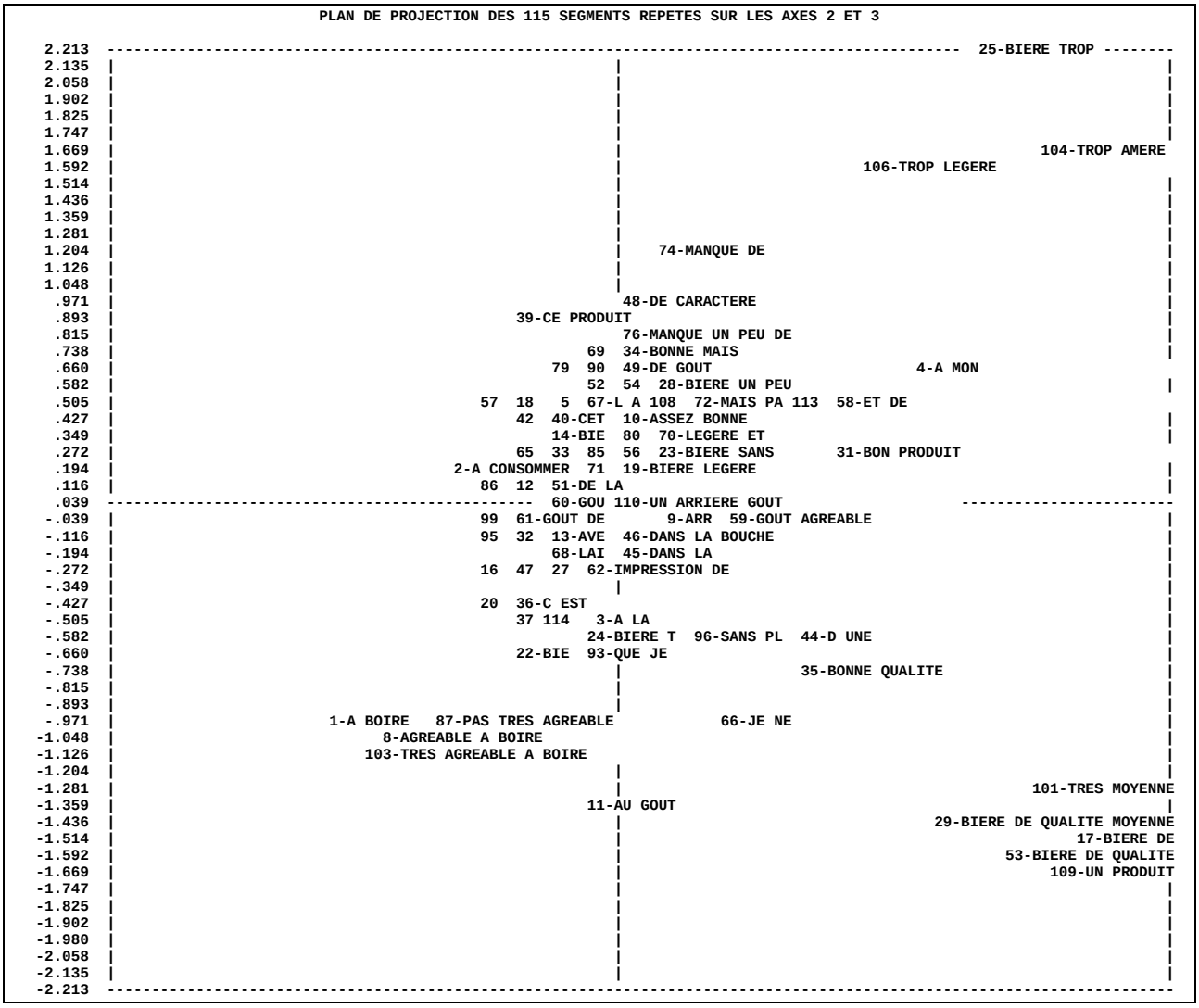


Figure 14 : Analyse factorielle des correspondances des segments répétés sur le tableau lexical entier

L'inertie du premier facteur (2,31%) est composée à 95% du segment *bon produit*.
Les axes 3 et 4 semblaient davantage informatifs, c'est pour cette raison qu'ils sont représentés sur la Figure 14.
D'après les segments de plus forte contribution sur ce plan, on peut reconstituer les phrases *bière trop amère* ou *bière trop légère* dans le coin supérieur droit, *bière de qualité moyenne* en bas à droite et *très agréable à boire* en bas à gauche.
Cette seconde analyse peut permettre d'affiner la première et de lever quelquefois des ambiguïtés au niveau du contexte immédiat d'un mot.
EXEMPLE : ***pas*** *très agréable* a une signification opposée à *très agréable*.

D'une façon globale, nous pouvons apercevoir des éléments intéressants à travers cet essai. Mais sur le plan pratique, le fait que les premiers facteurs n'utilisent qu'une faible part de l'information contribue à rendre difficile la synthèse visuelle. De plus, nous avons vu que les premiers facteurs peuvent être triviaux (inertie contribuée par une seule forme ou segment répété).
Rapidement, nous nous rendons compte que ce type d'analyse ne répond pas à nos attentes.

Analyse par regroupement

Nous avons choisi de réaliser l'analyse par regroupement suivant le croisement sexe x notes hédoniques. Elle est réalisée sur le tableau lexical agrégé comprenant en lignes les mots ou segments répétés issus des commentaires sur l'impression globale des six produits et en colonnes les classes de réponses regroupées suivant les modalités du croisement de variable suivant :

Tableau 20 : Modalités croisées sur la notation et le sexe

Modalités (libellé)	Effectifs
NoH1 = notes faibles (1, 2 et 3) / Homme	49
NoF2 = notes faibles / Femme	36
NoH3 = notes moyennes (4 et 5) / Homme	185
NoF4 = notes moyennes / Femme	124
NoH5 = notes élevées (6 et 7) / Homme	106
NoF6 = notes élevées / Femme	92

Les traitements réalisés sont une Analyse Factorielle des Correspondances, une classification ascendante hiérarchique et une recherche des éléments caractéristiques.

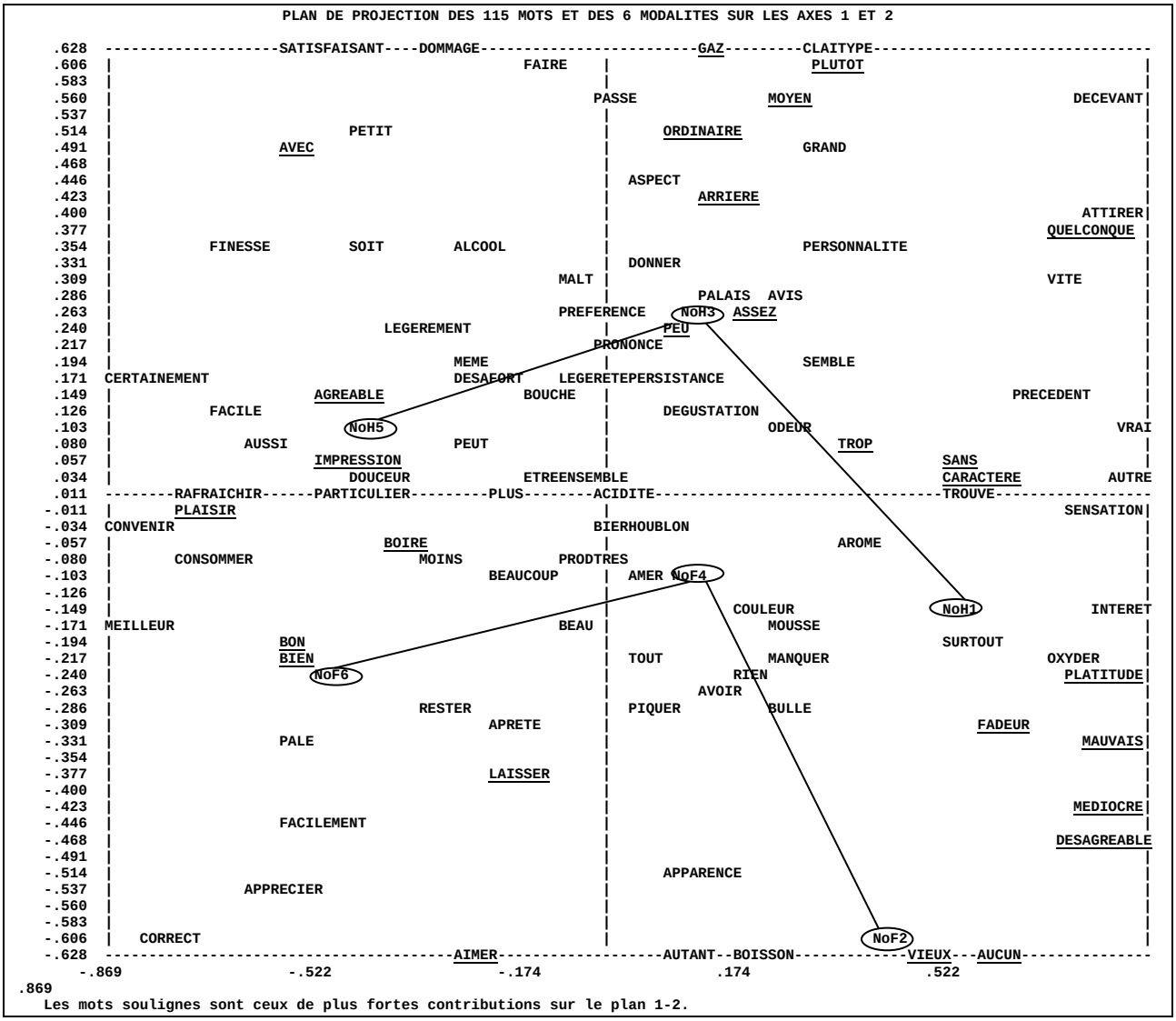


Figure 15 : Analyse factorielle des correspondances des mots sur le tableau lexical agrégé

La Figure 15 montre que l'axe 1 (39 % de l'inertie) est un axe de qualité : à droite, nous trouvons une abondance de termes négatifs et à gauche des termes positifs. Quant au second facteur (21 %), il permet de différencier les termes modérés (partie supérieure) des termes extrêmes (partie inférieure).
De même, la différence de comportement entre sexes mise en valeur d'une façon très nette. Pour une même classe de notes, les hommes ont des commentaires généralement moins sévères que ceux des femmes.

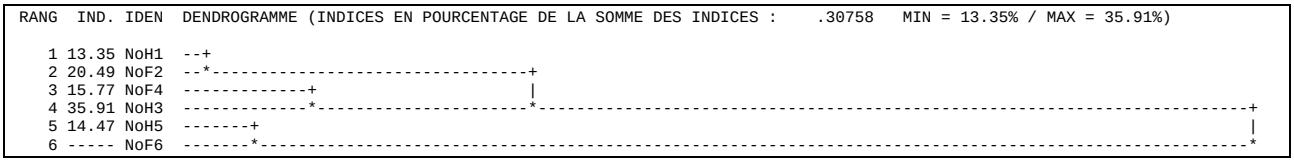


Figure 16 : Classification hiérarchique ascendante sur les mots

La classification d'après l'AFC sur les mots nous montre qu'il faut tout de même relativiser sur l'ampleur de ce phénomène. En effet, les inerties inter-classes de notes (donc les différences entre classes de notes) restent supérieures aux inerties inter-sexes.
Nous pourrions dire que, lors de ce regroupement, deux effets existent dans la façon dont les consommateurs expriment leurs impressions globales, l'effet notes et l'effet sexe; mais que l'effet note reste plus important que l'effet sexe.
L'analyse des segments répétés mène aux mêmes conclusions.

Tableau 21 : Mots, réponses et segments caractéristiques

Modalités	Mots	Réponses (+ pour les fréquences et * pour le Chi-2)	Segments
NOH1	désagréable fadeur bière quelconque très médiocre trop	+ trop fade. + assez désagréable. + bière trop fade, sans goût. * c'est une bière très désagréable en bouche qui laisse un goût amer. * bière de qualité inférieure, goût presque désagréable.	- bière trop - pas très - bière sans
NOF2	goût aucun manquer désagréable mauvais mousse	+ goût désagréable. + manque de goût. + fade, mauvais goût, juste buvable. * manque de goût, laisse de l'amertume. * mauvais produit, sans mousse, manque de caractère, produit très fade, plat.	- n'a - n'a pas - manque de - pas du tout
NOH3	moyen typé peu faire plutôt assez grand gaz arrière	+ moyen. + moyenne. + ordinaire. * bière assez agréable à boire, qualité moyenne. * bière pas très agréable à boire. Pas typée et avec trop de mousse; peu pétillante. * bière un peu légère.	- pas très agréable - un peu trop - un peu légère - goût assez - arrière goût
NOF4	arôme manquer vieux trop aimer amertume	+ trop amère. + trop légère, manque d'amertume. * un peu trop amère à mon goût. * bière assez bonne, mais qui manque un peu de goût.	- peu de - à mon goût - n'est - n'est pas
NOH5	agréable avec bon finesse certainement apprécier impression satisfaisant	+ très bonne impression. + bon produit. + bonne bière. * très bonne avec un goût marqué qui ne dénature pas cette bière agréable. * bière très agréable, légère.	- avec un - très bonne - un produit - goût agréable - très agréable
NOF6	bon boire correct bien facilement plaisir agréable	+ très bonne. + bonne bière. + bon produit. * bon produit agréable à boire. * bière agréable. Bonne. * agréable à boire, bon goût. Une impression de légèreté au palais. Finalement bonne bière qui me plait.	- bonne bière - bonne impression - agréable à - me plait - agréable à boire - bonne qualité - j'aime - est très

Les mots, réponses et segments caractéristiques visualisés sur le Tableau 21 permettent d'observer l'effet note mais il est moins évident de déceler l'effet sexe. Nous avons ici encore déterminé des groupes de consommateurs à partir des données nominatives et non par rapport aux éléments constitutifs du texte.

En conclusion, cette approche permet bien d'analyser les jugements caractéristiques des bières en fonction des différentes modalités. De 1992 à 1995, l'ensemble des études consommateur était traité de cette façon grâce à des macro commandes EXCEL rendant son utilisation rapide.

Mais elles comportent un inconvénient majeur qui est la perte du contenu du texte. En effet, nous reprendrons les lignes de G. TEIL dans sa thèse **[TEIL91]** pour dire que :
La liste des occurrences les plus fréquentes d'un texte ne permet pas de faire ressortir les grands thèmes des discours, leur contenu ou leur articulation. Les analyses d'occurrence produisent des représentations trop distantes du texte pour permettre une analyse fine de contenu.

2.2.2. L'analyse lexicale par contexte avec ALCESTE

Présentation de l'outil

ALCESTE est un logiciel d'analyse des données textuelles, issu du CNRS avec le soutien de l'ANVAR **[REIN86]**. Il a été adapté au MACINTOSH et au PC par la société IMAGE spécialisée en mathématiques appliquées et développement de logiciels scientifiques.

La méthode de classification de ce logiciel part du principe qu'il n'existe pas de savoir contextuel a priori lors d'un dépouillement d'enquête quand nous voulons faire une analyse de contenu. L'analyse se déroule donc sans connaître à l'avance le sens d'un corpus.
ALCESTE propose donc *de rendre compte de l'organisation interne d'un discours plutôt que de rendre compte de différences statistiques entre les divers textes d'un corpus* **[REIN93]**.

L'intérêt du TEPRAL pour cette approche découle d'un désir de comparer les performances de SPAD.T avec celles d'ALCESTE comme le CREDOC le décrit dans ces études (**[AUCO91]**, **[BEAU95]**, **[BEA93a]**, **[BEAU94]**, **[LAHL92]**, **[LAHL93]**, **[LION91]**, **[YVON90]**).

Découpage en unités de contexte

ALCESTE propose de segmenter arbitrairement le corpus de départ en Unités de contexte de façon à garder une certaine indépendance vis à vis des variations de segmentation (mots, expressions, conjonctions, phrases, paragraphes, …).
Nous distinguons les unités de contexte initiales ou U.C.I. des unités de contexte élémentaires ou U.C.E.
L'U.C.I. est la plus grande unité de contexte définissable sous ALCESTE. Elle représente un découpage " naturel " du corpus (les réponses à une question ouverte, les chapitres d'un livre ...)
L'U.C.E. est un segment de texte d'au plus 240 caractères, inclus dans une même U.C.I. et préférentiellement terminé par une ponctuation. Elle peut comprendre plusieurs phrases courtes ou un morceau de phrase longue.

L'ensemble du vocabulaire est réduit suivant deux procédures :

- à l'aide d'un dictionnaire des racines : c'est un algorithme qui reconnaît les mots outils pour les éliminer et les racines des principaux verbes irréguliers pour les réduire à leur forme infinitive.
- à l'aide d'un algorithme particulier pour traiter les formes non reconnues par le dictionnaire des racines. Celui-ci ne réduit une forme que dans la mesure où, d'une part, d'autres formes commençant par la même racine existent dans le corpus traité et, d'autre part, dans la mesure où les terminaisons de ces formes sont reconnues comme des suffixes valides retrouvés dans le dictionnaire des suffixes.

Calcul des tableaux de données

Nous retrouvons les U.C.E. en ligne et les formes réduites en colonne. L'intersection des lignes et des colonnes contient soit des 1 (présence de la forme dans l'U.C.E), soit des 0 (absence de la forme dans l'U.C.E).

Recherche des classes caractéristiques

Cette méthode de classification descendante hiérarchique a été mise au point pour d'une part répondre aux problèmes de traitement des grandes matrices creuses et d'autre part pour construire plus facilement des classes de cardinal élevé bien différenciées les unes des autres.

Cette méthode est une procédure itérative. Il s'agit de commencer à fixer le nombre de classes finales[22]. La première classe analysée comprend toutes les unités retenues. Ensuite à chaque pas, l'algorithme cherche la partition en deux de la plus grande des classes restantes en maximisant le critère du Chi2. Ceci prend fin lorsque le nombre d'itérations est épuisé.
Les formes réunies à la base de la hiérarchie sont "proches" en ce sens qu'elles sont dans le même environnement lexical.

Application

Présentation du corpus

1049 consommateurs ont dégusté en aveugle 6 bières de 6 marques différentes. Seules les réponses aux questions ouvertes sont utilisées. Nous avons choisi de nous limiter à celles concernant le goût. L'analyse portera donc sur 6160 réponses.

[22] Il existe également une option qui permet de sélectionner automatiquement la partition stable ayant le plus grand nombre de classes, afin que le programme puisse se poursuivre jusqu'à son terme sans intervention de l'utilisateur

Analyse statistique

Nous effectuons une Classification Descendante Hiérarchique sur une matrice binaire croisant en lignes les réponses des consommateurs et en colonne les formes réduites ou lemmatisées.
Nous retrouverons le principe de la méthode dans **[REIN83]** et **[REIN86]**.
Des contraintes de calcul ont été fixées au départ :

- moins de 15 classes demandées (10 demandées ici)
- tableau de données binaires de dimensions inférieures à 1400 x 10 000 avec au plus 60 000 chiffres 1.
- moins de 200 UCI demandées par classes

Résultats et interprétation

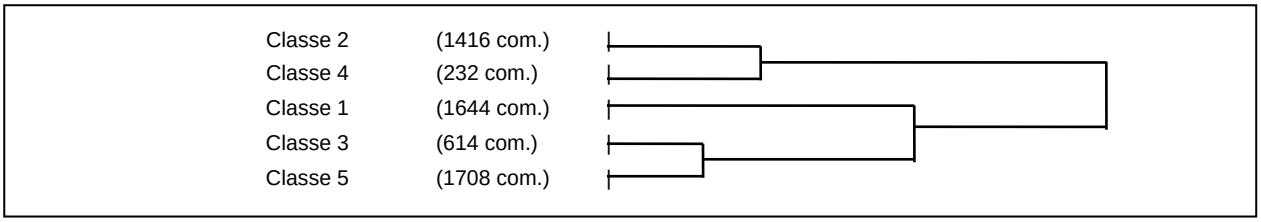


Figure 17 : Classification hiérarchique descendante des 6160 commentaires sur le goût de 6 produits différents

Le dendogramme de la figure 4 nous montre les résultats d'une classification descendante hiérarchique. Les six produits sont répartis en cinq classes décrites grâce au Tableau 22.

Tableau 22 : Description des cinq classes par les expressions, le vocabulaire spécifique et les réponses caractéristiques

Classe et produits spécifiques	Expression	Vocabulaire spécifique	Exemple de réponse caractéristique
1 Prod 1 Prod 2 Prod 3 Prod 5	belle couleur goût bière goût agréable en bouche bonne odeur	goût, bière, couleur, mousse, fade, belle, manque, blond, alcool, fort, class..., neutre, caractère, précédant, claire, faible, joli, normal.	Amertume assez prononcée. Couleur intermédiaire brune/blonde. Goût moyen. L'absence de mousse est décevant.
2 Prod 4 Prod 6	dans bouche mauvais goût on dire sensation désagréable.	désagréable, cidre, mauvais, acide, impression, eau, dégoût, bizarre, fait, vraiment, surprenant, difficile, étrange, indéfinissable	Jus de fruits, pomme, poire, ou autre fruit de saison été/automne, mais difficile à définir. Cela m'a rappelé mon enfance et le moment, dans les cidraies, de la fabrication du cidre.
3 Prod 5 Prod 3	dans bouche en bouche sur langue goût amer très agréable	bouche, laisser, palais, piquer, rester, langue, se, marque, picoter, gorge, agressif, lourd, passe, rape, durable, absorption, nuance, tranquille.	Bon. Glisse bien. Reste sur la langue et dans le palais.
4 Prod 6	envie en reprendre donner envie envie en reboire pas envie très agréable donner envie boire	envie, donner, , boire, reboire, reprendre, verre, finir, goûter, terminer, chaud, faire, acheter, chaleur, immédiatement, recracher, passer, horrible, déplaire.	Goût trop amer qui me donne des frissons. J'ai du mal a finir le verre. Elle donne envie de l'extérieur, mais dès qu'on la goûte elle désillusionne.
5 Prod 2 Prod 1	légèrement amer peu amer très agréable goût agréable agréable à boire bon goût	agréable, amer, rafraîchir, léger, fraîcheur, sucre, légèrement, fruité, désaltérer, doux, pétiller, frais, légèreté, parfum, douceur, bien-être.	Légèrement amer légèrement sucré et pétillant. Rafraîchissante désaltérante et légère.

La classe 1 semble plutôt associée à une perception globalement positive. Le vocabulaire est relativement descriptif : goût, odeur, couleur.
La classe 2 semble plutôt associée à une perception globale négative du goût, le vocabulaire utilisé étant de l'ordre de l'Impression.
La classe 3 semble aussi plutôt associée à une perception globalement positive. Le vocabulaire est plus " engagé sensuellement ", notamment c'est le contact " physique " qui est investi.
La classe 4 est plus ambivalente quant à l'attirance ou le rejet, l'amertume étant soit valorisée soit dévalorisée, ce qui laisse supposer pour le produit 6 un goût caractérisé. Le vocabulaire est davantage un vocabulaire d'action, de prise de position.
La classe 5 semble aussi plutôt associée à une perception globalement positive. Le vocabulaire investit davantage sur la dimension de la " soif ", insistant donc sur une impression globale du produit.

En conclusion, deux grands types de produits se différencient selon la valorisation globale positive ou négative.
Il est intéressant de noter une gamme de réactions différenciées chez les personnes testées allant de la prise de position active à une attitude davantage descriptive.
Le vocabulaire tourne ainsi autour de quatre pôles :

- le dégoût ou l'envie
- la saveur, l'odeur et la couleur
- le contact physique, corporel
- un sentiment plus global de fraîcheur ou de légèreté

La bonne différenciation des produits en fonction de ces pôles laisse supposer que leurs qualités propres engagent le sujet testé vers tel ou telle gamme d'évocations.

Cette approche met en évidence cinq types de jugements, caractérisant l'ensemble des six produits. Elle n'a pas été employée dans la même perspective que l'étude précédente réalisée avec SPAD.T puisque ici, l'ensemble des commentaires n'a pas été segmenté par produit. La comparaison serait plus évidente en recommençant le même traitement pour chaque produit. Néanmoins, nous obtenons globalement le même type de résultats qu'avec SPAD.T, ce dernier étant mieux adapté au traitement des questions ouvertes que ALCESTE.

2.2.3. Le réseau de mots associés avec CANDIDE™

Présentation de l'outil

Le programme CANDIDE™ a été conçu et mis au point au Centre de Sociologie de l'Innovation l'Ecole des Mines de Paris avec l'aide de l'INIST et du CNRS. Le principe de cette méthode est basé sur des calculs d'indice d'association et plus précisément sur l'algorithme du programme LEXIMAPPE™ qui a également été mis au point à l'Ecole des Mines de Paris **[MICH88]**.

L'intérêt pour cette approche a été influencé par l'équipe d'analyse sensorielle de l'INRA qui a travaillé avec G. TEIL sur le programme "Clavier organoleptique" pour le Ministère de l'Agriculture **[TEI94a]**.

Sélection des mots

Les mots représentatifs du corpus sont sélectionnés[23] au travers d'un dictionnaire interactif (l'utilisateur est libre de choisir les mots qu'il juge représentatif d'une phrase). C'est une étape d'indexation assez fastidieuse suivant la nature du corpus.

Classification

Le programme effectue ensuite une classification particulière qui consiste à comparer les mots deux à deux pour construire des classes assimilées à des thèmes. Chaque thème ne peut pas contenir plus de dix mots. Le(s) mot(s) dont les relations sont les plus représentatives du thème lui donne son nom.

L'algorithme consiste à définir l'association entre deux mots dans un corpus comme le produit des probabilités d'avoir un mot quand nous avons l'autre. C'est le coefficient d'équivalence vu à la section 2.1.3 (Formule 7).

L'ensemble des cooccurrences est trié par ordre de valeur d'indice d'équivalence décroissant.
Le programme prend un mot et les neufs premiers mots qui lui sont le plus fortement associés pour former un agrégat ou thème.
Chaque thème est orienté horizontalement par un indicateur de la position d'un thème au sein du réseau des thèmes appelé centralité (obtenu en multipliant la moyenne des liens externes par la proportion de thèmes associés au thème en question) et verticalement par un indicateur de la structure interne d'un thème appelé densité (obtenu en multipliant la moyenne des liens internes par la proportion de liens interne par rapport au nombre de liens internes possibles).
Sur le diagramme stratégique, la centralité a été remplacée[24] par la notion de fréquence (nombre de jugements ayant contribué à la construction d'un thème) car le découpage des thèmes n'est pas réellement fiable. En effet, un thème comportant plus de dix mots est découpé arbitrairement en deux. Il peut donc apparaître avec une centralité très forte due à un lien considéré abusivement comme externe très fort alors qu'il serait plutôt marginal.

[23] Cette étape n'est pas obligatoire : deux choix sont possibles suivant le type d'analyse : une indexation exhaustive ou au contraire, réduite
[24] C'est une spécificité de Candide™ par rapport à Leximappe™

Ces deux indicateurs permettent de positionner les thèmes dans un diagramme stratégique de la forme suivante :

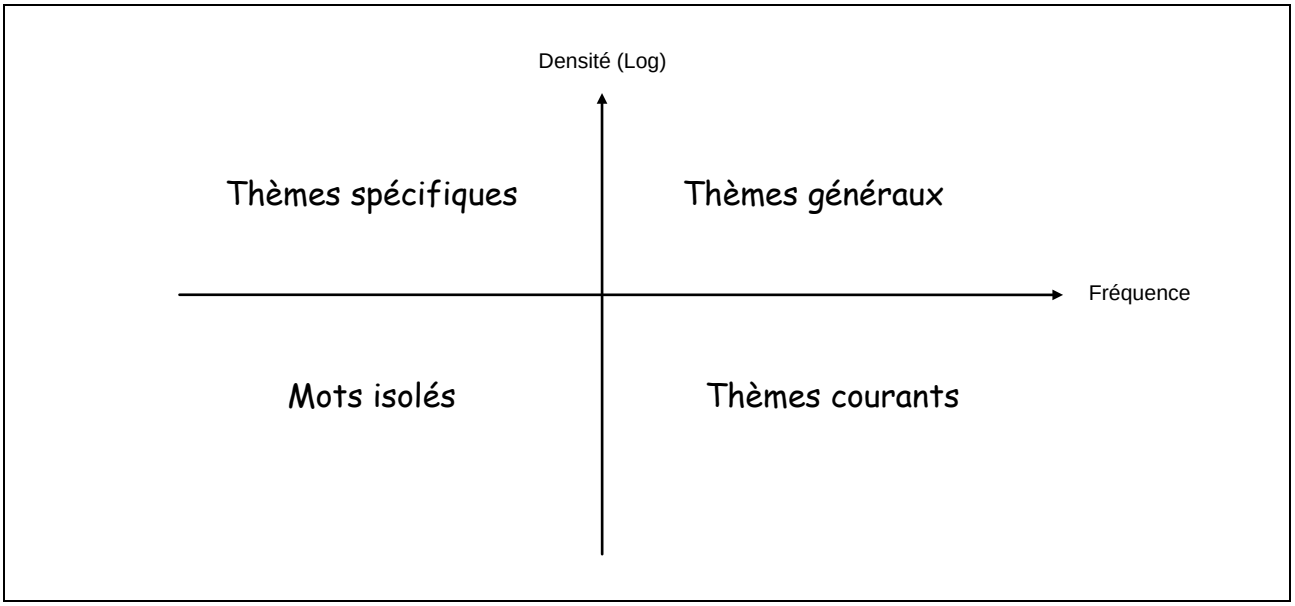


Figure 18 : Représentation générale d'un diagramme stratégique

Application

Présentation du corpus

600 consommateurs ont dégusté en aveugle 9 bières de 6 marques différentes. Seules les réponses aux questions ouvertes sont utilisées. Nous avons choisi de nous limiter à celles concernant les principales qualités de chaque produit. L'analyse portera donc sur 5400 réponses.

Analyse statistique

La construction du réseau des mots associés est établie sur une distance particulière, le coefficient d'équivalence sur l'ensemble des réponses.
Les produits ont été ajoutés en variable supplémentaire dans la construction des réseaux (ou thèmes).

Le réseau se compose de thèmes qui sont identifiés par le ou les mots les plus représentatifs de ce thème.
Par défaut les thèmes contiennent 10 mots, cependant ce paramètre peut être modifié.
Son but est plus d'étudier les grands thèmes sur lesquels les consommateurs s'expriment que de décrire chaque produit.

Résultats et interprétations

Un thème relativement important dans l'analyse de ce fichier sur les qualités est celui du **rafraîchissant.** Ce thème est assez stable et rassemble 17 % des jugements. Ceci est visible sur la figure suivante :

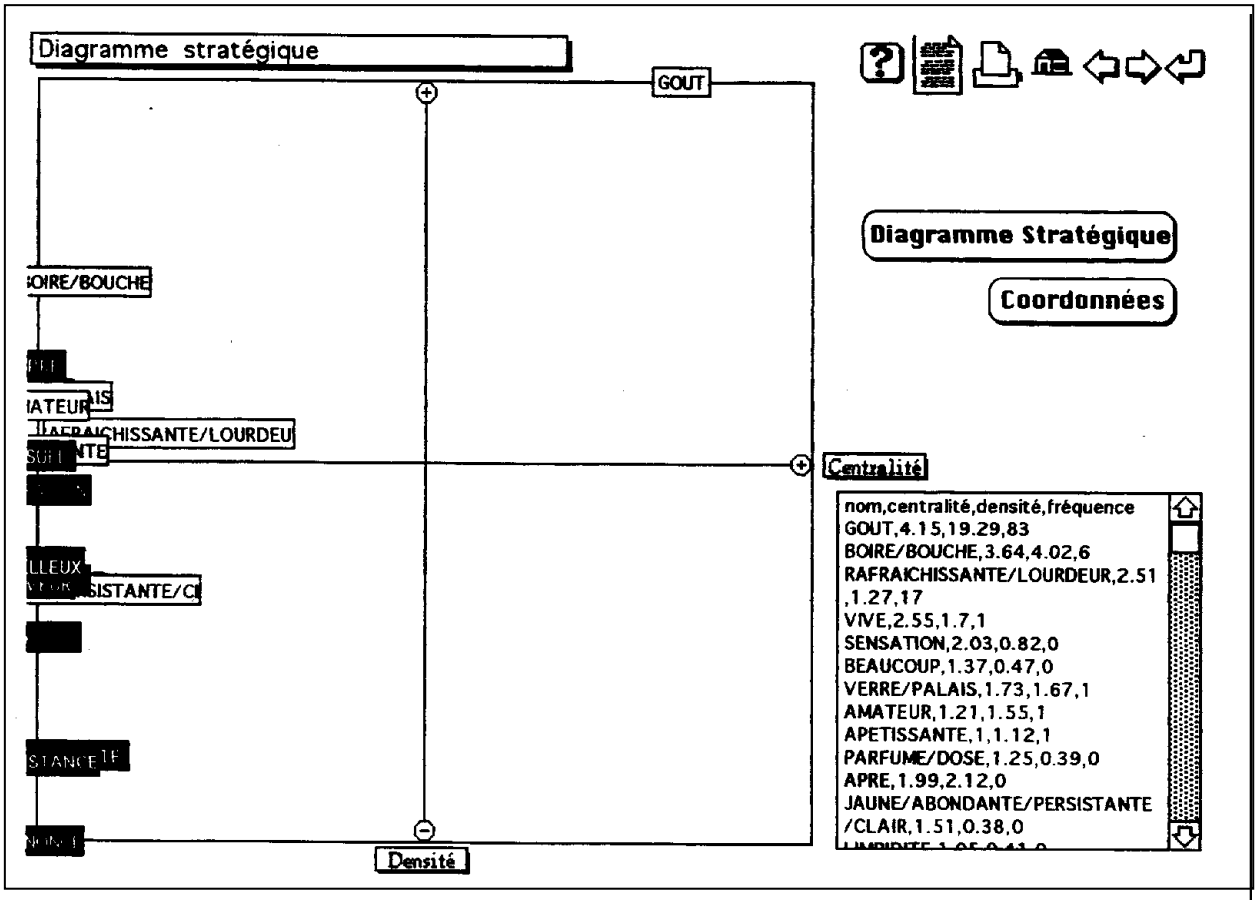


Figure 19 : Diagramme stratégique

Le diagramme stratégique montre aussi que le thème "goût" (concerne 83% des jugements) est très stable et commun à l'ensemble des commentaires, le thème "boire/bouche" est stable et rare, le thème "prononcé" est instable et rare.

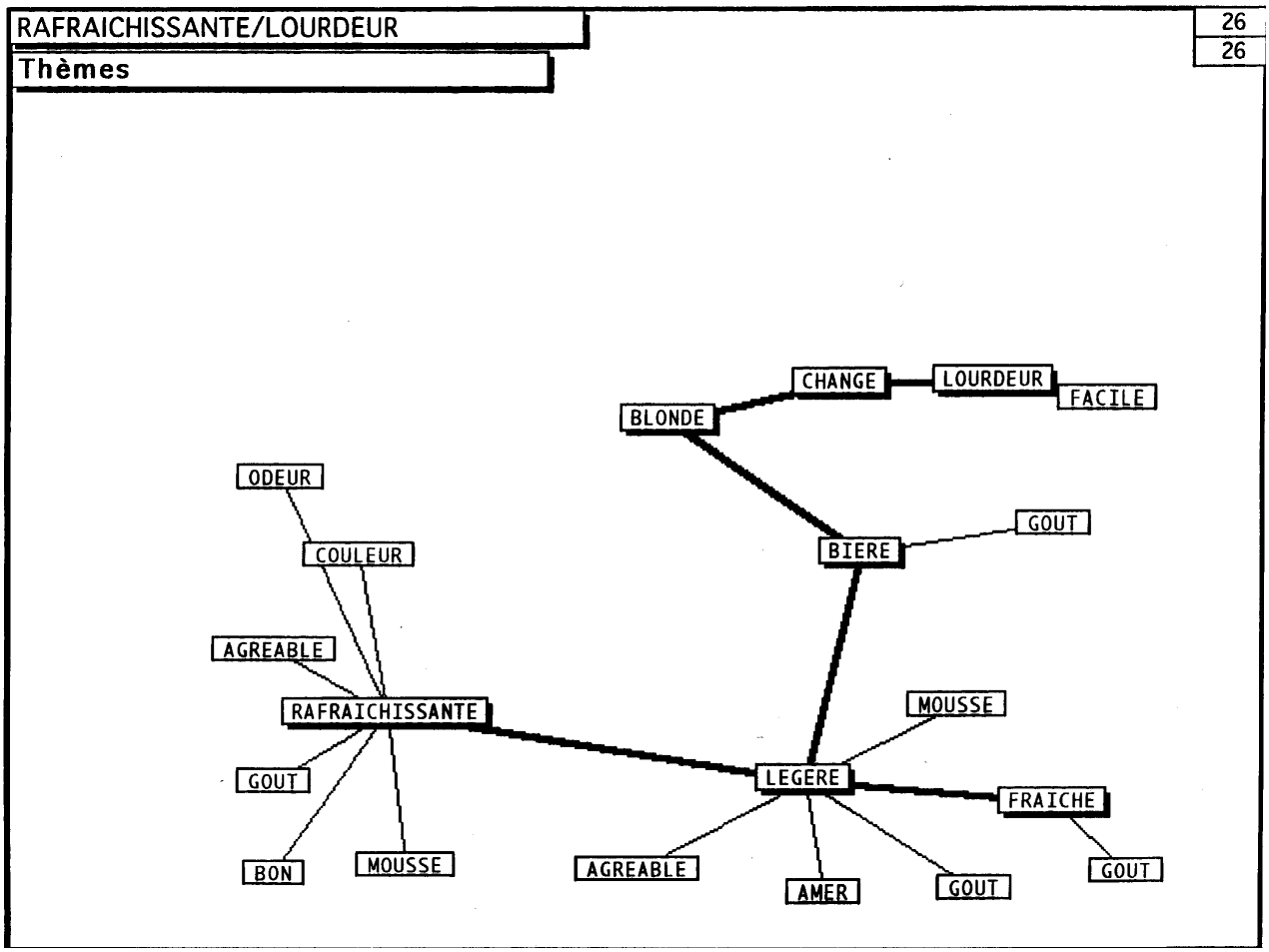


Figure 20 : Réseau des mots associés sur le thème du rafraîchissant

Les consommateurs attachent beaucoup d'importance au rafraîchissant d'une bière. C'est un mot cité très souvent en commentaire sur les qualités des produits.
Nous constatons que les gens associent souvent ce terme à la notion de légèreté (commentaire type : "*bière légère et rafraîchissante"*). En plus, lorsqu'ils parlent de légèreté, ils mentionnent également la fraîcheur du produit. ("*bière fraîche*", "*goût frais*", ...)
Nous observons également dans ce thème la notion de lourdeur. Le contexte de ce mot est très cohérent et systématiquement cité en terme d'absence ("*pas de lourdeur*") souvent complété de "*qui change des bières blondes classiques*" (d'où la liaison du mot change).
Enfin, le mot carbonique est présent sur le graphe car un consommateur l'a cité quatre fois en disant "*bière blonde peu carbonique*".
Nous remarquons les liaisons (traits fins) de rafraîchissant et légère avec plusieurs thèmes voisins.
Quand les consommateurs s'expriment sur ces termes, ils parlent donc également d'autres aspects du produit, à savoir : la mousse, la couleur, le goût, l'amertume.

Pour conclure sur cette application, nous pouvons dire que les résultats n'ont pas été très probants dans cette étude. Le diagramme stratégique montre peu de thèmes très représentatifs des commentaires de consommateurs. Ceci peut sans doute provenir du fait qu'il y a eu peu de traitement des données brutes au départ (phase de sélection des mots). Nous avons été aussi freinés par l'aspect « boîte noire » du module de classification. Enfin, l'application ne permet pas d'établir des comparaisons faciles entre deux produits. Néanmoins, la représentation en réseau nous semble très appropriée pour une interprétation simple et rapide des résultats.

Les trois approches (SPAD.T, ALCESTE et CANDIDE) que nous venons de détailler ont chacune apportée des propriétés intéressantes pour le traitement des commentaires libres de consommateurs. Durant ces quatre années, l'utilisation de ces méthodes a confirmé la richesse de l'information textuelle mais des lacunes techniques nous ont incités à aller plus loin dans nos investigations sur les méthodes de traitement.
En effet, elles ont déclenché une ouverture d'esprit sur la façon de segmenter les unités de décompte, d'effectuer un calcul statistique et de représenter les résultats. C'est une des conséquences de l'intérêt porté sur les recherches d'un laboratoire spécialisé en bibliométrie, le Centre de Recherche Rétrospectives de Marseille. Cet intérêt s'est concrétisé par cette thèse afin de réaliser un échange de connaissances entre les méthodes de bibliométrie et de traitement des commentaires de consommateurs.
Les premiers travaux ont consisté à éprouver les outils et méthodes du CRRM sur les données consommateurs. La section suivante en est donc la résultante.

2.2.4. le réseau de segments avec INFOTRANS, DATAVIEW et MATRISME

Présentation des outils

INFOTRANS est un logiciel de reformatage de références bibliographiques. Il est développé et commercialisé par *Information & Communication* à Freiburg[25] **[INFO94]**.
DATAVIEW est un logiciel de bibliométrie développé par le CRRM[26] **[ROST93]**.
MATRISME a été mis au point grâce à la collaboration entre LEPONT[27] et le CRRM. Il est capable de représenter graphiquement le contenu d'une matrice de fréquences sous la forme d'un réseau **[BOUT96]**.

La méthodologie mise en œuvre est une approche d'analyse de textes pratiquée en bibliométrie. La bibliométrie est plus particulièrement axée sur l'exploitation de corpus de textes représentant de références bibliométriques **[ROST96]**. Les principes de cette discipline sont de dégager à partir d'un grand volume de notices bibliographiques les tendances générales de leurs contenus et d'offrir une grille de lecture en déterminant les structures sous-jacentes à ces données.

[25] Information & Communication, Alte Str.66, D-79249 Freiburg Merzhausen, Allemagne
[26] CRRM, Centre scientifique de St Jérôme 13397 MARSEILLE CEDEX 20
[27] Laboratoire Le Pont, Université de Toulon et du Var IUT TC, BP 132 83957 LA GARDE CEDEX

Codage

Lors de ces enquêtes consommateurs, l'acquisition des commentaires se réalise sous forme papier (commentaires saisis de façon manuscrite sur des formulaires, voir section 3.1). Des opératrices les saisissent ensuite électroniquement par lecture et décryptage des formulaires (voir section 4.1.4). Le mode même de cette acquisition de données engendre deux problèmes majeurs :

- un grand nombre de termes erronés : les fautes de frappe systématiques ou occasionnelles, les fautes d'orthographes, les erreurs de lecture, les fautes de français.
- une très grande hétérogénéité du vocabulaire et des expressions employés (voir section 1.1.3).

Une telle diversité de termes impose un traitement préalable de correction des erreurs et de codage des commentaires (voir section 4.2) pour réduire le vocabulaire et augmenter la signification des traitements statistiques ultérieurs. Cette démarche correspond tout à fait au principe statistique de la bibliométrie qui au détriment d'une perte d'information offre un gain de signification.

Ce codage passe par plusieurs étapes :

- corrections des erreurs répertoriées
- élimination des mots outils
- repérage des locutions et liaison des termes qui les composent
- lemmatisation
- regroupement synonymique
- gestion des ambiguïtés (polysémie et homographie)
- marquage des termes spécifiques à l'analyse sensorielle
- homogénéisation des termes de quantification

Les cinq premières étapes sont totalement automatisées grâce à l'établissement de lexiques spécifiques au produit alimentaire étudié (lexique des erreurs, des mots outils, des locutions, des lemmes, des synonymes). Ces lexiques sont systématiquement appliqués aux données brutes grâce au logiciel de reformatage INFOTRANS (nous l'aborderons à la section 4.2). Un tel logiciel ne sait pas traiter les aspects de catégorisation grammaticale et de syntaxe de phrase. Seuls des traitements de reconnaissance et de manipulation de formes graphiques sont réalisables.
L'automatisation complète de la sixième étape nécessiterait une analyse morpho-syntaxique et sémantique impossible à envisager avec un reformateur. Elle n'est donc que semi-automatisée. Un lexique des termes potentiellement ambigus a été établi. Ce lexique permet de les " marquer " de façon à pouvoir les retrouver facilement en fin de traitement. Il faut alors lire le contexte pour évaluer par quel autre terme il doit être remplacé (une table des termes ambigus et de leurs remplaçants potentiels a été rédigée pour aider le correcteur). Cette étape correspond au précodage de C. MULLER dans **[MUL92a]**.

Enfin la dernière étape est là encore basée sur l'emploi de lexiques. Cette fois-ci, non pour réduire le vocabulaire mais uniquement pour " marquer " les mots ou locutions très appréciés pour l'analyse sensorielle. Ainsi, 6 catégories sont construites : les termes faisant appel à l'*arôme*, au caractère *hédonique*, à la *perception*, à la *saveur*, à la *texture* et à l'*aspect*.
Tous les termes appartenant à ces classes étant marqués (voir exemple ci-dessous), il devient plus facile de les manipuler pour construire les tableaux croisant les termes des différentes catégories (voir Tableau 23 et ANNEXE 10).

EXEMPLE :
Avant postcodage
GOUT AGREABLE. ARRIERE GOUT ASSEZ AMER MAIS NE SUIT EN RIEN LA QUALITE DU PRODUIT. TRES RAFRAICHISSANT

Après postcodage
*@GOUT *AGREABLE. @ARRIERE_GOUT ASSEZ µAMER. QUALITE TRES_FAIBLE BIERE. TRES RAFRAICHISSANT*

Tableau 23 : Sigles des différentes catégories sensorielles représentées dans le vocabulaire des consommateurs

Sigle	Catégorie
@	arôme
*	hédonique
|	perception
£	saveur
§	texture
&	aspect

Tous ces lexiques sont bien évidemment remis à jour après analyse de chaque nouveau corpus de commentaires libres. Chaque étude apportant son lot de nouvelles fautes, de nouvelles expressions, de nouveaux synonymes, il est indispensable de les prendre en compte pour les traitements futurs. Ce système de codage est donc conçu dans un contexte évolutif.

Segmentation et comptage des associations de segments

Les données obtenues après codage offrent plusieurs voies de segmentations. La première est de tout simplement considérer toutes séquences de caractères encadrés d'un espace ou d'un point comme étant des formes graphiques à dénombrer. Un problème se pose alors lorsqu'il faut comptabiliser les associations de formes graphiques. Il faut rappeler que l'objectif de ces études d'analyse sensorielle est de cartographier au plus juste chaque produit testé.

Pour cela, non seulement la liste des sensations évoquées par les consommateurs est importante, mais encore plus les associations de sensations. Or dans le cas où l'unité statistique textuelle serait celle indiquée ci-dessus, deux cas de comptage d'association sont envisageables.

Cas A : associations des termes intra-phrase
Seuls les termes appartenant aux mêmes phrases se retrouvent associés. Pour l'exemple présenté plus haut, les associations seront :

*@GOUT ↔ *AGREABLE*
@ARRIERE_GOUT ↔ ASSEZ
@ARRIERE_GOUT ↔ µAMER
ASSEZ ↔ µAMER
QUALITE ↔ TRES_FAIBLE
QUALITE ↔ BIERE...

Dans ce cas, les associations précisant que le consommateur a trouvé le produit *agréable* avec un *arrière-goût amer* ou *très rafraîchissant* avec un *arrière-goût amer* sont négligés. Or ce sont justement ce type d'associations qui paraissent les plus intéressantes.

Cas B : associations des termes intra et inter phrase
Pour essayer de récupérer les associations précédentes, il est possible alors de considérer tous les couples de termes intra et inter phrases. Ce comptage fait bien ressortir les associations omises précédemment comme :

**AGREABLE ↔ RAFRAICHISSANT*
**AGREABLE ↔, µAMER*

mais il prend aussi en compte des associations comme :

**AGREABLE ↔ @ARRIERE_GOUT*
@ARRIERE_GOUT ↔ TRES_FAIBLE

voire :

**AGREABLE ↔ TRES*

ou :

µAMER ↔ TRES

Ces dernières associations sont indésirables et ne peuvent être prises en compte lors de l'analyse de la cartographie des associations.

C'est pour cela qu'une troisième solution a été envisagée. Puisque les phrases dans les commentaires libres sont le plus souvent très concises et que la phase de codage a réduit leur composition aux idées essentielles, nous pouvons considérer ces phrases comme des entités très homogènes, comme des concentrés d'information. L'unité statistique élémentaire peut alors être ramenée à l'échelle de la phrase. La segmentation pour le dénombrement de ces unités se fait donc grâce au point. Les associations des segments obtenus sont comptabilisées uniquement à l'intérieur d'un commentaire libre. Dans notre exemple, ce traitement donne les associations suivantes :

*@GOUT *AGREABLE ↔ @ARRIERE_GOUT ASSEZ μAMER*

*@GOUT *AGREABLE ↔ QUALITE TRES_FAIBLE BIERE*

*@GOUT *AGREABLE ↔ TRES RAFRAICHISSANT*

@ARRIERE_GOUT ASSEZ μAMER ↔ QUALITE TRES_FAIBLE BIERE

@ARRIERE_GOUT ASSEZ μAMER↔ TRES RAFRAICHISSANT

QUALITE TRES_FAIBLE BIERE ↔ TRES RAFRAICHISSANT

Cartographie des associations de segments

Le dénombrement des fréquences d'apparitions des segments (phrases codées) ainsi que le dénombrement des fréquences des co-présences des couples de segments sont des processus totalement automatisés grâce à exploitation du logiciel bibliométrique DATAVIEW **[ROST93]**. Parmi bien d'autres types de résultats, ce logiciel permet de ré-exprimer ces comptages sous la forme d'un tableau symétrique distribuant en ligne et en colonne l'ensemble des segments présents dans les corpus analysés. Une cellule d'un tel tableau comporte dans la diagonale, la fréquence d'apparition d'un segment, et hors de la diagonal, la fréquence des co-présences d'un couple de segments.
Le tableau obtenu est alors exporté vers le logiciel MATRISME spécialisé dans la génération automatique de réseaux **[BOUT96]**. Il produit une représentation infographique du contenu du tableau sous la forme d'un réseau (voir Figure 21). Les segments du corpus sont symbolisés par les nœuds du réseau tandis que les arcs reliant les nœuds représentent la fréquence de co-apparition des couples de segments. Contrairement aux analyses d'inertie, la position des nœuds les uns par rapports aux autres ne dépend pas d'une métrique mesurant les distances. Ces positions sont fonction d'une mesure d'évaluation de l'esthétisme du graphe obtenu (optimiser l'espace occupé, réduire le nombre d'intersections, interdire les chevauchements de nœuds, limiter la longueur des arcs). Seules les nuances de couleur (ou d'épaisseur) des arcs donnent des indications sur les intensités d'association entre les nœuds (les segments).

Un simple clic sur un nœud du réseau permet à l'utilisateur de remonter aux commentaires comportant le segment correspondant. Une organisation hypertextuelle des commentaires favorise ensuite une navigation de l'utilisateur en fonction des segments communs.

Application

Présentation du corpus

1038 consommateurs ont dégusté en aveugle 6 bières de marques différentes. Seules les réponses aux questions ouvertes sont utilisées. Il s'agit des réponses à la question "Décrivez l'ensemble des sensations que vous a procuré ce produit". L'analyse portera donc sur 6228 réponses.

Analyse statistique

L'ensemble des segments issus du codage des commentaires libres de consommateurs est exprimé sous la forme d'une matrice carrée et symétrique à partir du logiciel DATAVIEW. Cette dernière comporte l'ensemble des segments en première ligne et en première colonne, chaque cellule C_{ij} correspond à la fréquence d'apparition de la paire de segments i-j.
Cette matrice est le point d'entrée pour construire un graphe à l'aide du logiciel MATRISME. A partir de là, deux techniques complémentaires peuvent être utilisées pour rechercher les segments représentatifs :

Le réseau peut être réalisé sur l'ensemble des segments quelles que soient leurs fréquences. Nous obtenons alors une représentation illisible en forme de « pelote de laine ». En effet, parmi les segments, figurent des associations rares et leur prise en compte nuisent à la cohérence du graphe. Il est donc nécessaire de procéder à un seuillage afin d'éliminer un certain nombre de liens.
MATRISME permet en effet de paramétrer non seulement les fréquences des paires de segments pour éliminer le bruit ou l'information triviale mais aussi le nombre d'associations entre deux segments pour rendre le graphe plus lisible (voir Figure 21).

Résultats et interprétations

Pour un produit donné, nous obtenons le réseau de segments visible sur la Figure 21 obtenus à partir de 1038 commentaires libres. Seuls les segments ayant une fréquence supérieure à 4 et les relations supérieures à 1 sont représentés sur ce réseau. La valeur présente à côté du segment correspond à sa fréquence. Les trois nuances graphiques des fréquences des co-présences de segments sont expliquées par la petite fenêtre intitulée *Légende des arcs.*
D'une façon générale, nous remarquons deux notions centrales dans le réseau. Il s'agit de rafraîchissant et agréable qui sont fortement cités et fortement associés à des notions annexes telles que le désaltérant, la fraîcheur, la légèreté et la douceur.
Ces observations rejoignent tout à fait celles faites sur le test avec le logiciel CANDIDE™, du moins avec la notion du rafraîchissant.

Nous remarquons qu'il y a peu de segments dans la représentation du réseau. Ils apparaissent plutôt de façon isolée. Alors que le réseau est davantage composé de mot isolé.

L'examen des commentaires codés nous montre que les segments sont plus fréquents à l'intérieur des grandes phrases tandis que les mots isolés ressortent davantage dans les phrases courtes.
Or, les commentaires libres de consommateurs sont composés en grande majorité de phrases courtes. C'est donc pour cette raison que le réseau de mots isolés prédomine sur la Figure 21.

Il est intéressant de voir que cette approche permet non seulement de perdre le moins possible d'information mais surtout de retracer les associations de sensations "complexes" sous la forme des segments codés énoncés par les individus.
Enfin, avec cette approche nous pouvons créer une représentation par produit. Nous pouvons envisager d'effectuer des comparaisons avec plusieurs produits en cherchant les segments communs à tous les graphes.

Cependant, une information cruciale nous manque : la nuance quantitative et/ou hédonique cité avec le segment ou la forme isolée. En effet, dans plusieurs cas les ambiguïtés subsistent : le terme *désaltérant* a-t-il été cité en tant que *très désaltérant*, *pas désaltérant* ou encore *moyennement désaltérant* ? Cette idée qui nous semble spécifique au sensoriel a été analysée et une solution est proposée plus loin à la section 4.2.2).

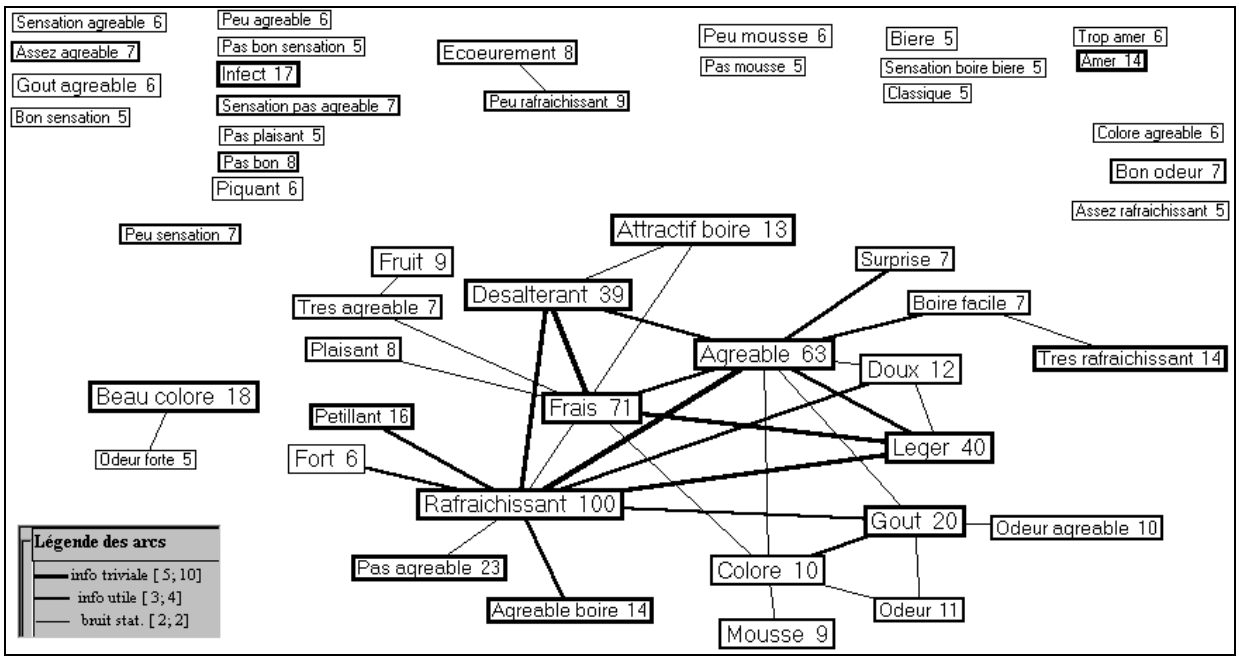


Figure 21 : Réseau des segments des commentaires libres sur les sensations que procure la bière Y

2.3. Bilan sur les méthodes utilisées

Le Tableau 24 va nous permettre de faire une synthèse sur les quatre types de méthodes que nous venons d'exposer concrètement.

Tableau 24 : Avantages et inconvénients des méthodes utilisées

Logiciel	Avantages	Inconvénients
SPAD.T	× Traitements statistiques usuels et surtout socio-démographiques × Accès au contexte de citation (mais de façon indirecte) × Traitement statistique assez rapide	× Pas facile à utiliser × Moyennement convivial (retour aux données initiales) × Significatif sur un grand échantillon (valeur test) × **Pas de diminution de la dispersion du vocabulaire** (lemmatisation) × Pas de levée d'ambiguïté lexicale automatique × Difficile à interpréter
ALCESTE	× Accepte les corpus de grande taille × **Diminution de la dispersion du vocabulaire** (Lemmatisation) × Récupération des données (ASCII) × Classification stable × Facile à interpréter	× **Pas de levée d'ambiguïté lexicale automatique** × Pas facile à utiliser × Pas rapide × Pas convivial × Pas d'accès au contexte de citation × Adapté aux textes littéraires et aux discours
CANDIDE™	× Accepte les corpus de grande taille × Facile à utiliser	× **Pas de levée d'ambiguïté lexicale automatique** × Pas d'accès au contexte de citation × Par rapide × Classification non homogène × **Pas de comparaison de graphe** × **Pas facile à interpréter sans l'aide d'un spécialiste**
INFOTRANS / DATAVIEW / MATRISME	× Accepte les corpus de grande taille × **Diminution de la dispersion du vocabulaire** × Facile à utiliser × Accès au contexte de citation × **Comparaison de graphe**	× Pas de levée d'ambiguïté lexicale automatique × **Pas facile à interpréter sans l'aide d'un spécialiste**

D'une manière générale et d'après le tableau ci-dessus, nous pouvons remarquer que chacune des trois méthodes amène des avantages et des inconvénients différents.

L'expérience avec SPAD.T a permis de mettre en évidence le fait qu'il soit indispensable de réduire la diversité du vocabulaire. En effet, ce logiciel ne permet pas de valoriser la richesse des commentaires libres des consommateurs **[MART93]**.

Avec ALCESTE, nous avons d'une part, confirmé la nécessité de lemmatiser le vocabulaire et d'autre part, montré qu'il était important de prendre en considération l'ambiguïté lexicale (voir la section 4.2.1) et l'ambiguïté sensorielle (vu à la section 1.1.5) en ayant accès au contexte de citation.

CANDIDE™ a confirmé la nécessité de lever les ambiguïtés lexicales et sensorielles et nous a donné l'exemple d'un mode de représentation très simple à interpréter.

Nous n'oublierons pas de noter que SPAD.T et ALCESTE proposent des méthodes statistiques incontestables (caractérisation des variables, stabilité de la classification hiérarchique descendante). Par contre pour CANDIDE™, le type de classification hiérarchique à lien simple basé sur l'indice d'équivalence a l'avantage d'être simple mais présente l'inconvénient de ne pas être homogène et représentative de la structure réelle d'un corpus.
Ce système de découpage ne permet pas d'obtenir des groupes cohérents et c'est une source d'erreur dans l'interprétation. En effet, la première classe qui sera construite contiendra l'information triviale et la dernière classe contiendra tous les termes qui n'auront pas été agrégés.

ALCESTE et SPAD.T ont souvent été comparé sur les questions ouvertes **[LAHL93]**. Les résultats ont conclu à la constatation suivante : STAD. T est particulièrement adapté à des textes courts, très redondants avec une structure grammaticale pauvre, tandis que ALCESTE est plutôt orienté vers l'étude des textes longs et littéraires.

Enfin, l'approche de réduction du vocabulaire avec INFOTRANS est un bon compromis pour obtenir un gain de signification statistique suffisant d'une façon simple et automatique. Les différentes phases de codage paraissent relativement bien au point pour permettre de répondre à une grande partie de cet objectif. Une analyse par catégorisation et une analyse sémantique pourrait nettement faciliter la phase de traitement des ambiguïtés mais ces approches seraient très coûteuses lors de leur mise au point. La technique choisie est peut-être frustre linguistiquement mais elle offre l'avantage d'être accessible à tous et rapide à mettre en place.
L'approche de segmentation des textes en phrase codée paraît bien appropriée au type de données collectées dans le cadre d'études d'analyse sensorielle (concision des phrases composées). Elle reste tout de même a être confortée lors d'études ultérieures.
L'analyse des associations et des dépendances d'idées par la représentation cartographique sous forme de réseau avec DATAVIEW et MATRISME est particulièrement bien adaptée à la phase d'interprétation. L'interprétation d'une étude ne peut s'envisager sans le soutien des professionnels du domaine étudié (dans notre cas les professionnels en analyse sensorielle), il est préférable que les supports d'analyse ou de communication soit le plus accessible possible. La représentation en réseau a cet avantage d'être compréhensible par tous sans aucun apprentissage spécifique, ce qui n'est pas le cas des méthodes basées sur une construction mathématique relativement complexe et difficilement explicable aux non-initiés **[ROST98]**.

La possibilité de retourner aux commentaires originaux d'un segment par simple " clique " sur le nœud du réseau lui correspondant, puis de naviguer dans l'ensemble des commentaires par liens hypertextes en fonction des segments auxquels il est associé, offre un outil d'aide à l'interprétation et de validation incontestable. Cette fonctionnalité de génération automatique de fichier hypertexte à partir des commentaires originaux, structuré selon les associations exprimées dans le tableau analysé, est un atout supplémentaire. Pouvoir passer de la représentation synthétique du réseau aux données brutes qui ont permis de la construire est à nos yeux un instrument indispensable à la bonne réussite d'une telle analyse de contenu de textes.

2.4. Evaluation d'outils

Les études décrites dans le 2.2 nous ont permis de manipuler les commentaires libres de consommateurs. Le bilan (voir 2.3) est loin d'être négatif puisque nous savons maintenant quelles seront les phases de traitement sur lesquelles nous devrons particulièrement nous pencher : le codage et le mode de représentation.
Mais avant de commencer à mettre en œuvre une nouvelle solution de traitement, nous avons cherché à vérifier si des outils existants étaient capables de résoudre ces problèmes.

Le travail a consisté à évaluer l'ensemble des outils capables de réaliser une ou plusieurs phases de notre chaîne de traitement vu à la section 1.2.2 sur lesquelles les méthodes utilisées jusqu'à présent comportent des lacunes, à savoir :

- la collecte des commentaires libres,
- le codage,
- le traitement statistique
- la représentation graphique

Cela signifie donc que nous n'avons pas hésité à augmenter notre champ de vision au niveau des domaines d'application.
Les outils[28] retrouvés dans le tableau qui suit ont tous été évalués avec le même jeu de données. Ils sont décrits suivant cinq colonnes :

Produit	Informations générales sur le logiciel : son nom, le système d'exploitation sur lequel il a été testé et sa version
Approche	Discipline dans laquelle le logiciel est utilisé : analyse linguistique, analyse lexicale, analyse de contenu, analyse de discours, recherche documentaire, traitement d'enquêtes, intelligence artificielle, bibliométrie
Rôle	Rôle du logiciel dans la chaîne de traitement : collecte, codage, traitement statistique et représentation graphique

[28] On retrouvera en ANNEXES 4 et 5 les caractéristiques commerciales de ces produits et de ceux qui ont été observés sans être évalués

Commentaires Liste des avantages et des inconvénients :
saisie par ordinateur (internet, borne interactive), acquisition des données par OCR, récupération des données, questions ouvertes, convivial, facile à utiliser, rapide, diminution de la dispersion du vocabulaire, accès au contexte de citation, levée d'ambiguïté lexicale automatique, valeur de l'analyse statistique (comptage/tri, méthodes multidimensionnelles, classifications, réseaux de connexion …)

Tableau 25 : Description des logiciels

Produit	Discipline	Rôle	Description	Commentaires
FU MS DOS Vers. 3.56 (1990)	Recherche documentaire	Codage	Fu propose un reformatage de texte par ligne de commande. Le programme est inspiré du langage AWK (sous Unix).	**Avantages** × Diminution de la dispersion du vocabulaire × Accès au contexte de citation × Récupération des données (ASCII) **Inconvénients** × Pas de levée d'ambiguïté lexicale automatique × Pas convivial × Pas facile à utiliser × Pas rapide × Pas d'accès au contexte de citation
Lexico Macintosh 1994	Lexicométrie	Codage Traitement statistique	Ensemble de programmes lexicométriques composé de cinq modules. A partir d'un fichier texte, en s'appuyant sur la liste des caractères délimiteurs, le premier module opère la segmentation automatique du texte et calcule les occurrences des formes graphiques. Le programme crée ensuite une base de données numérisée. Le module de documentation permet de retrouver l'ensemble des contextes d'une forme sélectionnée par l'utilisateur. Le corpus peut être découpé en parties qui servent à construire le tableau lexical à partir duquel nous pouvons réaliser différentes analyses statistiques. D'autre part, les segments répétés sont calculés. Un module effectue une analyse des correspondances des tableaux (formes X parties) ou (formes X segments X parties). Enfin, un module calcule les accroissements spécifiques et les spécificités chronologiques.	**Avantages** × Accès au contexte de citation(concordances) × Facile à utiliser × Assez convivial **Inconvénients** × Limité en taille × Pas rapide × Pas de diminution de la dispersion du vocabulaire × Pas de levée d'ambiguïté lexicale automatique × Statistiques classiques

SNR MS DOS Vers. 1.5 (1988)	Recherche documentaire	Codage	SNR propose un reformatage de texte par des table de transfert.	**Avantages** × Diminution de la dispersion du vocabulaire × Récupération des données (ASCII) **Inconvénients** × Pas de levée d'ambiguïté lexicale automatique × Pas convivial × Pas facile à utiliser × Limité en taille × Pas d'accès au contexte de citation
Le Sphinx Lexica Windows 1997	Traitement d'enquêtes	Collecte Codage Traitement statistique Représentation graphique	Il traite toutes les étapes d'une enquête du questionnaire au rapport : Conception du questionnaire récupérable avec sa mise en page dans un traitement de texte, saisie optimisée des réponses, échantillonnage et redressement, dépouillement automatique (tri à plat, tableaux croisés ...) à partir du module de base Le Sphinx Primo. Le Sphinx Plus est un module supplémentaire de statistique avec lequel nous pouvons réaliser des analyses factorielles, des classifications, des typologies, des analyses de variances à partir des enquêtes construites avec Le Sphinx Primo ou en important des données de bases externes. Enfin, Le Sphinx Lexica est un module qui propose d'analyser le texte libre en réalisant de l'analyse de contenu assistée par ordinateur, de l'analyse morpho-syntaxique, des mesures de spécificité, des calculs de segments répétés.	**Avantages** × Saisie sur ordinateur et sur Internet × Traitement des questions ouvertes × Accès au contexte de citation × Diminution de la dispersion du vocabulaire × Assez convivial × Facile à utiliser × Rapide × Levée d'ambiguïté lexicale automatique partielle (SYLEX, voir ci-dessous) **Inconvénients** × Récupération des données via des modules supplémentaires × Statistiques classiques

Sylex Unix et Windows 1995	Analyse linguistique	Codage	Ces composants d'analyse linguistique proposent de réduire la dispersion lexicale dans les réponses à des questionnaires ouverts et peuvent conduire à une aide à la codification des réponses. Les fonctionnalités sont divisées en deux pôles d'intérêt, la lemmatisation et l'analyse syntaxique complète d'un texte : Lemmatisation : analyse lexicale et syntaxique d'une phrase, désambiguïsation et production de la liste des formes canoniques des mots de cette phrase ainsi que de leur catégorie syntaxique. Dans ce module, sont intégrés les locutions, les mots composés, les structures verbales incluant les pronoms antéposés, postposés, les passés composés et surcomposés, les adverbes, les formes négatives, interrogatives et passives. Analyse complète : production des liens syntaxiques par le traitement des structures déclaratives, interrogatives, négatives, des relatives, des compléments de nom, des fonctions sujet, COD, COI, de certains circonstants et d'une partie des conjonctions notamment dans les syntagmes nominaux. Résolution de certaines références pronominales, possessives et démonstratives.	**Avantages** × Diminution de la dispersion du vocabulaire × Accepte les corpus de grande taille × Levée d'ambiguïté lexicale automatique partielle **Inconvénients** × Peu convivial × Pas facile d'accès × Pas d'accès au contexte de citation
Technociel Macintosh 1995	Analyse linguistique	Codage	L'analyse comporte 3 phases : 1. le texte est analysé linguistiquement de façon automatique. Il s'agit de découper l'ensemble du texte en unités cohérentes sur le plan syntaxique 2. les unités extraites sont ensuite regroupées (automatiquement) en fonction de leur occurrence et de leur contenu. Apparaissent ainsi les premiers groupes de mots importants 3. Les groupes de mots les plus importants sont alors affinés et triés selon des méthodes semi automatiques. Cette phase donne naissance à des listes de segments répétés.	**Avantages** × Découpage automatique × Accès au contexte de citation **Inconvénients** × Pas rapide × Pas de diminution de la dispersion du vocabulaire × Données textuelles non exploitables directement par un logiciel statistique classique × Pas de levée d'ambiguïté lexicale automatique

TEWAT Unix et PC 1995	Bibliométrie	Traitement statistique Représentation graphique	Il repose sur l'algorithme d'analyse relationnelle des données. Il s'agit d'identifier une relation de similitude entre les documents et les descripteurs qu'ils contiennent (calculé sur l'indice…). La classification est réalisée selon le critère de Condorcet. Nous pouvons également effectuer une mesure de similitude de l'ensemble des relations entre les classes (inertie inter-classes).	**Avantages** × Accepte les corpus de grande taille × Ne fixe pas à priori le nombre de classes × convivial × Modules de traitements statistiques intéressants (classification) × Résultats assez faciles à interpréter × Traitement rapide × Accès au contexte de citation **Inconvénients** × Pas facile d'utilisation (importation des données) × Pas de comparaison de classifications × Très cher × Paramétrage très long

Text Navigator Unix 1997	Analyse linguistique	Codage Traitement statistique Représentation graphique	Il se divise en deux fonctionnalités : l'analyse textuelle et la classification. Cette dernière découle de TEWAT. L'analyse textuelle comporte quatre phases : × Le pré-traitement (ou segmentation du corpus) qui identifie les parties à indexer, × l'étiquetage grammatical permet d'affecter à chaque mot sa catégorie grammaticale (chaînes de Markov) × la phase de repérage d'expression caractéristique permet de lemmatiser et de regrouper les locutions × l'indexation qui sélectionne statistiquement les unités qui seront utilisés comme descripteurs (élimination des unités très et faiblement fréquentes).	**Avantages** × Levée d'ambiguïté lexicale automatique × Accepte les corpus de grande taille × Diminution de la dispersion du vocabulaire × Accès au contexte de citation × Traitement assez rapide **Inconvénients** × Lexiques électroniques spécialisés à rajouter × La phase de pré-traitement n'est pas standardisée × Pas facile à utiliser × Pas de comparaison de classifications × Très cher × Paramétrage très long

Tri-Deux MS DOS Vers. 2.2. (1995)	Analyse de contenu Analyse synaptique Lexicométrie	Codage Traitement statistique Représentation graphique	Logiciel de dépouillement d'enquêtes utilisant des méthodes factorielles et post-factorielles et comportant 16 modules. Il permet de faire de l'analyse de contenu.	**Avantages** × Laisse les traits pertinents émaner des données elles-mêmes × Récupération des données × Traitement des questions ouvertes × Modules de traitements statistiques intéressants (tris, multidimensionnelles, classifications, réseaux de connexion) **Inconvénients** × Pas convivial × Pas facile × Pas rapide × Pas de diminution de la dispersion du vocabulaire × Pas de levée d'ambiguïté lexicale automatique × Pas d'accès au contexte de citation × Résultats difficiles à interpréter

En 1996 aucun outil n'a été choisi car bien qu'apportant tous certaines spécificités intéressantes, ils se révèlent être soit trop pointus, soit trop généraliste ou encore trop coûteux.
Cependant, l'ensemble des méthodes et des outils qui ont été étudiés dans cette partie apporte tous des résultats intéressants.
Le principal atout est qu'ils nous ont aidés à mieux connaître les données textuelles et leur façon de les manipuler. Nous avons pu préciser toutes les opportunités de traitements liés aux données textuelles pour déterminer l'approche idéale pour des données sensorielles.
Ceci nous a permis de définir nos objectifs méthodologiques pour mieux les aborder.

3. Objectifs méthodologiques du traitement des commentaires libres de consommateurs

D'une manière générale, le traitement des commentaires libres de consommateurs va dépendre fortement des besoins que nous nous fixons pour utiliser au mieux ces données. Les objectifs en matière de résultats devront donc être d'abord défini pour déterminer le type de traitement le plus approprié. Aussi avons-nous privilégié les axes suivants :

- amélioration du mode de collecte
- amélioration du codage
- simplification de la lecture des résultats

Ces axes ressemblent de près à ceux que nous avions énoncés à la section 2.1. Ils sont en effet très liés au côté classique de la démarche de traitement sous forme de chaîne (vu à la section 1.2.2).
D'après l'étude des résultats provenant de travaux ultérieurs, nous sommes en mesure de dégager les avantages et inconvénients pour résoudre notre problématique.

3.1. Améliorer le mode de collecte

Cette étape englobe autant le mode de questionnement que le mode de récolte des données consommateurs.
Pour le premier, il s'agit de trouver la formule idéale de la question pour permettre à l'ensemble des consommateurs de comprendre ce que le questionnaire attend d'eux afin de transcrire le plus fidèlement leurs perceptions.
Le second devra permettre d'améliorer le système de gestion des tests consommateurs pour réduire et fiabiliser la saisie mais aussi pour motiver le consommateur en rendant le questionnaire interactif.
L'avancée de nos recherches sur cette problématique n'étant pas suffisamment conséquente, nous avons préféré les exposer dans la partie sur les perspectives dans la section 4.2. du chapitre III.

3.2. Améliorer le codage

La nature de nos données et l'état actuel des connaissances nous poussent à envisager un prétraitement des commentaires libres de consommateurs. Cette étape devra nous permettre de résoudre le problème paradoxal de la diminution de la dispersion du vocabulaire pour un gain en signification statistique.

3.2.1. Réduire la dispersion du vocabulaire

Le vocabulaire contenu dans les commentaires libres de consommateurs peut être caractérisé d'hétérogène : les mots vides se mélangent avec les mots pleins, les mots pleins sont déclinés sous différentes formes fléchies, synonymes ou antonymes, …
Ce premier constat nous pousse à envisager un tri suivi d'un codage approprié de façon à :

- regrouper les mots d'une même famille sous une même forme,
- regrouper les synonymes,
- éliminer les mots vides

Mais très vite, nous nous rendons compte que certains termes nous laissent dans l'embarras car ils peuvent renfermer soit la même signification pour deux orthographes différentes soit la même forme pour deux définitions différentes. Il s'agit pour la première des notions complexes qui engendrent une ambiguïté sensorielle auprès des consommateurs alors que la seconde relève de l'ambiguïté lexicale.

Les travaux de lexicométrie, nous l'avons vu à la section 2.1.2, ont déjà abordé ce genre de problèmes. Ils pourront nous servir d'aide dans la construction d'une norme spécifique de codage pour lemmatiser de façon automatique les commentaires libres de consommateurs **[LABB92]**.

3.2.2. Accéder au contexte de citation

Les travaux de N. MARTIN et M. ROGEAUX ont abordé l'intérêt d'étudier non seulement les descripteurs employés par les consommateurs mais aussi leur contexte de citation **[MART94]**. En effet, les commentaires contiennent trois types d'informations intéressantes et complémentaires pour l'analyse sensorielle :

- la description des sensations perçues
- l'intensité de ces sensations
- l'intensité du plaisir ressenti

Par exemple, lorsqu'un consommateur indique que le produit est peu fruité, il est intéressant d'apprendre qu'il parle du fruité, mais il est tout aussi important de savoir dans quelle proportion, surtout si elle a été énoncée !
De la même façon, le consommateur est spontanément de donner l'intensité du plaisir qu'il a ressenti en dégustant un produit. Cette information accompagne et complète non seulement les descripteurs mais aussi l'information sur l'intensité des sensations.

Prenons l'exemple précédent, lorsque le consommateur qualifie le produit comme peu fruité, est-ce que cela signifie pour sa préférence : assez ou pas assez fruité ?

Au contraire, G. TEIL ne s'intéresse qu'aux descripteurs en supprimant du lexique la plupart des mots mesurant un des critères retenus **[TEI94a]**. Nous notons cependant plusieurs termes conservés dans son analyse qui représentent pour nous des termes de quantification : fort, prononcé, léger, …

La difficulté réside dans la liaison du bon terme de quantification avec le bon descripteur. Il est d'autre part nécessaire d'organiser l'ensemble de ces termes d'intensité (sensation et plaisir).
Cette approche n'a jamais été abordée en partie pour ces raisons techniques. Pourtant il s'agit là d'un critère de distinction intéressant entre les données textuelles classiques et les commentaires libres de consommateurs après dégustation de produit alimentaire.

3.3. Faciliter la lecture des résultats

L'objectif de cette étape consiste à communiquer les résultats issus du traitement des données. Pour qu'elle soit réussie, la représentation synthétique doit être simple, claire et doit apporter une plus-value. Il est donc prudent de réfléchir à la façon de mettre en valeur l'information sans la biaiser.
La démarche consistera à représenter d'une façon synthétique le discours des consommateurs en perdant le minimum d'information. L'idée générale est de pouvoir représenter à la fois les termes descriptifs les plus cités et le contexte dans lequel ils ont été employés.
Les expériences avec les logiciels CANDIDE™ et MATRISME nous incitent à préférer la représentation en réseau (voir sections 2.2.3 et 2.2.4).
Cependant, nous souhaitons avoir la possibilité de comparer les graphes entre eux. Cela sous-entend la mise en place d'un repère selon lequel, les éléments du graphique pourront être répartis.

Les trois objectifs que nous venons d'énoncer sont spécifiques à notre approche sur le traitement des commentaires libres de consommateurs.
Or, ils correspondent à des méthodologies connues et employées en sciences de l'information (notamment en bibliométrie) ainsi qu'en lexicométrie. En effet, l'application de la bibliométrie en veille technologique aborde fréquemment les techniques automatiques de codage et de représentation des données textuelles. De même en lexicométrie, la réduction de la dispersion du vocabulaire est très utilisée pour analyser les discours ou les œuvres littéraires.
Nous nous en sommes inspirés pour mettre en place une nouvelle chaîne de traitement.

4. Mise au point d'une nouvelle méthode

L’analyse de l’existant à travers la bibliographie (voir section 2.1) et l’expérience DANONE (voir section 2.2) nous ont aidées à mieux cerner la démarche que nous devrons aborder pour valoriser les commentaires de consommateurs.
Cette dernière va donc consister à proposer une méthode de traitement entièrement paramétrable et modulable inspirées des méthodes lexicométriques et bibliométriques.
La Figure 22 schématise la chaîne de traitement composée de quatre étapes :

- la collecte
- le codage
- le traitement statistique
- la représentation graphique

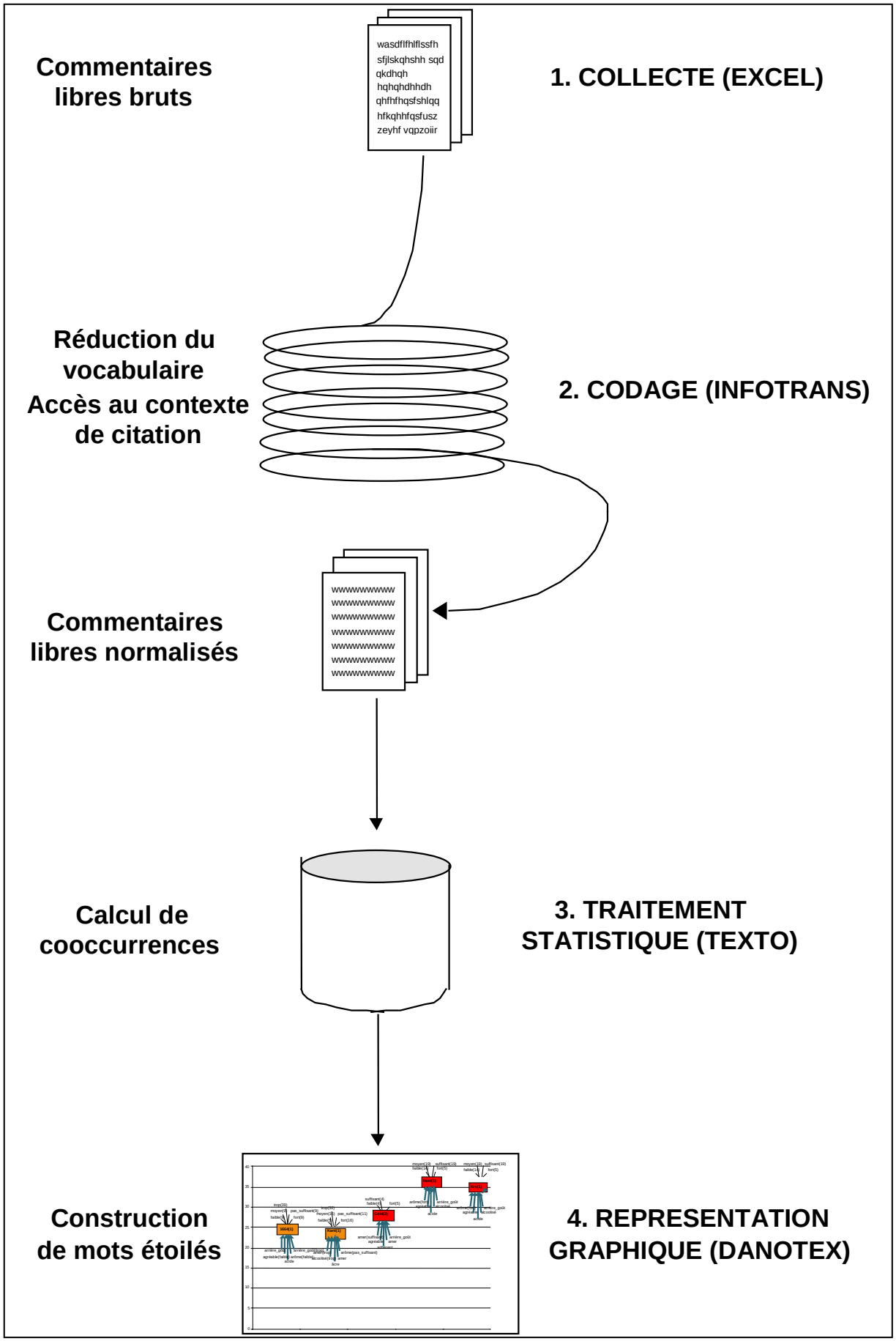


Figure 22 : Chaîne de traitement des commentaires libres de consommateurs

4.1. Collecte des commentaires libres

4.1.1. Caractéristiques d'un test consommateur au TEPRAL

Classiquement dans un test consommateur, l'échantillon représentatif est déterminé en fonction de la cible des enquêtés recherchée (méthode des quotas).
Les produits sont présentés aux consommateurs en mode monadique séquentiel. Pratiquement, le consommateur teste un premier produit, répond simultanément au questionnaire. Le deuxième produit est testé de la même façon, c'est-à-dire de façon indépendante vis à vis du premier.
La distribution des produits est réalisée selon un plan d'expérience (voir exemple en ANNEXE 1) de façon à équilibrer les séquences de dégustation.
Enfin, ce sont souvent des enquêtes en mode auto-administré[29] puisque les consommateurs remplissent eux-mêmes leurs questionnaires en dégustant le produit alimentaire.

4.1.2. Type de questionnaire

Questionnaire classique

Il est le plus souvent réalisé pour les tests consommateurs utilisés en cartographie des préférences (voir section 3.5. du chapitre I).
Il s'agit d'un questionnaire sur papier ou micro-ordinateur sur lequel nous retrouvons des questions classiques de renseignements sur l'état civil et sur les habitudes de consommation. Le consommateur doit également remplir une batterie de notes de préférences au fur et à mesure qu'il évalue les produits. Pour chaque produit, une ou deux questions ouvertes sont associées aux préférences (voir ANNEXE 3).

Questionnaire spécifique

Il a été mis en place uniquement pour collecter les commentaires libres de consommateurs (voir ANNEXE 2). Il comporte trois parties :

- une partie pour recueillir les commentaires libres des consommateurs sur leurs sensations après la dégustation de la première bière, ses qualités et ses défauts, les circonstances dans lesquelles elle serait consommée.
- une partie pour recueillir des informations personnelles.
- une partie identique à la première partie pour recueillir les commentaires libres des consommateurs sur la seconde bière.

Pour que l'enquêté ne soit pas tenter de comparer les deux produits, le mode opératoire est expliqué en début de test de manière écrite sur la feuille de questionnaire et de manière orale par l'enquêteur.

[29] L'enquêté se charge lui-même d'écrire la réponse sur papier, micro-ordinateur ou minitel

4.1.3. Libellé des questions

Les premiers questionnaires comportaient une seule question très générale, demandant aux enquêtés quelles étaient leurs impressions sur le produit qu'ils venaient de déguster. Ce mode de questionnement fut vite abandonné car l'information recueillie était très hétérogène. En effet, des éléments positifs et négatifs se retrouvaient mélangés.
La deuxième tentative fut de demander plus précisément ce que les consommateurs pensaient du goût, de l'aspect, l'odeur, … Malheureusement, les résultats montrèrent un grand nombre de répétition dans les réponses. Malgré les différentes questions, les consommateurs mélangeaient encore les types d'information, peut être à cause d'un problème de compréhension.
C'est alors que trois types de questions simples à interpréter ont été définis afin de collecter à chaque réponse des informations le plus homogène possible et non redondantes avec la réponse suivante ou précédente :

- *Vous venez de boire le produit X, quelles sensations vous procure-t-elle ?*
- *Citez les principales qualités de ce produit :*
- *Citez les principaux défauts de ce produit :*

Le consommateur y répond par écrit.
D'une manière indirecte, ce questionnaire permet au consommateur de faire une pause naturelle de quelques minutes entre la dégustation de deux produits.

4.1.4. Norme de saisie

La Figure 22 et les sections précédentes indiquent que les commentaires libres sont récoltés au moyen de questionnaires papiers. Ces derniers sont ensuite habituellement saisis à la main par une opératrice dans le logiciel tableur MICROSOFT EXCEL. En effet, les traitements des données textuelles se faisaient auparavant avec SPAD.T et étaient préparés par une macro-commande sous EXCEL.
Pour réaliser un traitement informatique des commentaires libres qui produise des résultats fiables et intéressant, deux conditions doivent être remplies. D'une part, la saisie du texte doit obéir à des règles rigoureuses et, d'autre part, ils doivent subir une opération de codage préalable au traitement statistique (**[LABB90]**, **[LABB92]**, **[MUL92a]**).
En ce qui concerne le premier point, le principe de base part du principe de limiter les contraintes spécifiques sur la saisie.
Aussi, avons nous mis en place une norme de saisie simple des commentaires libres afin de limiter les risques d'erreurs et de préparer les données au codage (voir en ANNEXE 11).

4.2. Codage

" *L'enregistrement d'un texte n'est pas fait pour interpréter, coder les sens, les contenus ou les thèmes, analyser les liens et fonctions grammaticales et réunir les flexions sous les lemmes, mais pour fournir simplement au chercheur un matériel identique ou presque à l'édition de référence.* " **[LAFO85]**

La stratégie de recherche sur le traitement des commentaires libres de consommateurs a été choisie en fonction d'objectifs bien définis. Rappelons que nous avons d'une part choisi de récolter les commentaires libres de consommateurs pour recueillir le vocabulaire des consommateurs et avoir accès aux réponses spontanées qui décrivent mieux les sensations. D'autre part, la méthode de traitement doit conserver le minimum de termes pour un maximum de signification avec la possibilité de conserver la nuance d'intensité avec laquelle chaque mot a été cité.

L'étape que nous appelons codage va nous permettre d'accéder à ces objectifs en proposant une " norme " de dépouillement automatique des commentaires libres de consommateurs. Ainsi, il sera plus aisé d'établir des dénombrements sur des unités bien définies et normalisées.

Cette étape a été réalisée à l'aide de INFOTRANS présenté ultérieurement à la section 2.2.4. Il s'agit d'un logiciel de reformatage de références bibliographiques (voir section 2.1.2). Il est donc le plus souvent utilisé dans le domaine documentaire.
Grâce à des listes de termes et des tables de transfert, il est capable de reformater un corpus de taille variable. Il réalise donc un cherche/remplace multiple mais peut également modifier la structure des phrases (en particulier, le rattachement des termes de pondération aux termes de description).
Ce logiciel a été choisi parce qu'il est simple d'accès et permet de réaliser une lemmatisation automatique. Les tables de transfert sont paramétrables grâce à un métalangage simple. Il s'intègre facilement dans une chaîne de traitement et peu communiquer avec la plupart des logiciels.

Notre codage va s'effectuer en deux temps :

- la première étape effectue un codage partiel et sélectionne les éléments ambigus qui pourraient avoir plusieurs significations. C'est le précodage.

- la deuxième étape consiste en un codage total des commentaires dont nous avons levé les ambiguïtés lexicales.

Chaque étape est inclue dans une logique de chaîne de traitement. Elles doivent être réaliser dans l'ordre car chacune dépend de l'autre (voir section 1.2.2).

4.2.1. Précodage

C'est la première étape de la chaîne de traitement. Elle comporte les phases suivantes :

- la lemmatisation
- le regroupement synonymique
- une partie du regroupement des locutions
- le repérage des termes ambigus (lexicalement et sensoriellement)

Lemmatisation

Nous nous sommes inspirés des travaux énoncés en section 2.1.2 pour déterminer l'ordre de priorité de la lemmatisation :

1. ramener les formes fléchies à l'adjectif ou aux participes masculins/singuliers,
2. ramener les formes fléchies au nom masculin/singulier,
3. ramener les formes fléchies au verbe à l'infinitif,
4. ramener les formes fléchies à la forme canonique.

Cela revient à dire que la forme de l'adjectif masculin singulier remplacera la forme nominale, verbale ou adverbiale. Nous avons en effet remarqué que l'adjectif est davantage approprié à la description des perceptions des consommateurs.

Ainsi, par exemple les mots *astringents*, *astringentes*, *astringente*, *astringent* et *astringence* seront regroupés sous la forme *astringent.*

Sur le plan technique, c'est une opération délicate qui suit deux étapes :

- la reconnaissance des formes
- la constitution d'un dictionnaire de lemmes

L'ensemble des lemmes est contenu dans une liste d'autorité représentant l'équivalent d'un dictionnaire des formes fléchies. Ce dernier peut donc être réutilisé à chaque codage et mis à jour en fonction des besoins.

Regroupements synonymique et antonymique

Une fois lemmatisés certaines formes ont une signification voisine. Ces synonymes sont donc regroupés sous une seule dénomination après la validation par des spécialistes[30].
Ainsi, par exemple les mots *rugueux*, *rude* et *râpeux* sont regroupés sous le mot *âpre*.

D'autre part, plusieurs formes antonymiques sont observées. La même opération de regroupement permet de retenir uniquement la forme positive.
Ainsi, par exemple les mots *moche*, *inesthétique*, *horrible* et *affreux* sont regroupés sous le mot *pas beau*.

[30] spécialistes en analyse sensorielle et en connaissance du consommateur

Comme pour les lemmes, les synonymes et les antonymes sont contenus dans une liste d'autorité.

Regroupement des locutions

Les formes contenues dans le corpus de commentaires libres proviennent quelque fois de locutions ou syntagmes. Ce sont des groupes de mots en séquence qui forment une unité de sens minimale à l'intérieur de la phrase (voir section 2.1.2).
Leur sens est lié à leur structure composée mais la segmentation des unités basée sur un découpage des formes découpe de la même façon les unitermes et les multitermes (voir section 1.2.1). Il est donc nécessaire de les considérer comme une seule forme en faisant intervenir un caractère de liaison " _ " pour regrouper les différents éléments de la locution.

Ainsi, par exemple les expressions *arrière-goût* et *bière sans alcool* sont regroupées sous les mots *arrière_goût* et *bière_sans_alcool*.

Levée d'ambiguïté lexicale

L'ambiguïté lexicale provoque une incapacité d'identifier clairement le concept désigné par un mot surtout lorsque nous travaillons sur des formes hors contexte. Il est donc impératif de lever l'ambiguïté qui touche certaines formes. Pour cela, il est nécessaire de faire appel à une analyse grammaticale.

Nous avons choisi de constituer un lexique du domaine dans lequel nous pouvons prévoir les cas classiques de polysémie et d'effectuer une analyse grammaticale intellectuelle **[MUL92a]**.
Les mots ambigus sont reconnus par le caractère "§". Grâce au tableau des termes ambigus (en ANNEXE 7), nous sommes capables de remplacer chaque terme polysémique par un terme monosémique.

Ainsi, par exemple le mot *doux* qui peut vouloir dire aussi bien sucré que faible dans les commentaires libres de consommateurs sera désigné sous la forme *§doux*. Après validation par des spécialistes[30] qui ont la responsabilité de lire le contexte de chaque mot ambigu, cette forme sera remplacée manuellement soit par *sucré* soit par *faible*.

4.2.2. Codage complet

La deuxième étape de la chaîne de traitement comporte plusieurs phases dont :

- la seconde partie du regroupement des locutions[31]
- l'élimination des mots vides
- l'homogénéisation des termes de quantification
- la pondération des termes de description

[31] suivant la même démarche qu'au précodage

Elimination des mots vides

Dans une phrase, certains mots sont plus chargés de sens sur le plan syntaxique que d'autres. Ils sont indispensables pour que la phrase soit cohérente mais perdent leur intérêt après une segmentation.
L'ensemble de ces éléments est usuellement appelé "mots vides" ou encore "mots outils" et ils sont représentés par les termes de liaisons, les articles, …
Ces derniers sont extraits et éliminés car ils sont assimilés au bruit ou encore à de l'information non pertinente (voir section 1.1.2).

Homogénéisation des termes de quantification

Le vocabulaire brut des consommateurs ressemble beaucoup à du langage parlé (voir section 1.1.3). Un grand nombre de nuances de quantification diverses sur la forme mais identiques sur le fond sont employées.

Or, l'analyse sensorielle a l'habitude de travailler à l'aide des échelles de notation. Nous avons donc dans un premier temps, regroupé l'ensemble des termes de quantification cités par les consommateurs au cours de divers sondages.
Puis, dans un second temps, ces termes ont été classés suivant les niveaux de quantification auxquels ils se rapportaient le mieux.
Deux types de nuances ont été distingués (ANNEXES 8 et 9) :

- les nuances qui quantifient les perceptions (très, moyen, faible)
- les nuances qui quantifient le plaisir (trop, suffisant, pas suffisant)

De plus, nous avons effectué une concordance de chacune des formes afin d'évaluer l'ensemble des nuances de cette forme et d'éviter les écueils liés au sens.
EXEMPLE :

Forme	**Nuances de la forme**
TRES	TRES BIEN
	TRES TRES
	VRAIMENT TRES
	TRES PEU
	PAS TRES
	TRES FAIBLE …

Ainsi, un grand nombre de nuances de termes de quantification et de jugement a pu être relevé à partir des commentaires libres de consommateurs.
Enfin, nous sommes parvenus à faire correspondre chaque nuance à un degré bien précis d'une échelle de valeur. Deux échelles ont été construites et validées par des spécialistes[30] (voir ANNEXES 8 et 9) :

- une échelle à sept niveaux
- une échelle à cinq niveaux

La première a été écartée car elle n'apporte pas plus d'information que l'échelle à cinq niveaux. Pour des raisons de simplicité, le traitement courant utilise donc essentiellement la dernière.

Ainsi, par exemple les expressions *un peu marqué* et *vraiment pas* sont respectivement regroupées sous les formes *assez_fort* et *très_faible*.

Pondération des termes de description

Dans la mesure où un terme descriptif est employé avec un certain degré de quantification, il est nécessaire qu'ils soient regroupés dans la même unité de décompte. En effet, comme pour les locutions, nous avons recherché à rattacher la nuance de quantification à son terme de description. Ici, les parenthèses remplaceront les caractères soulignés.

Ainsi, par exemple les expressions pas *assez de goût*, *pas d'alcool* et *trop amer* sont respectivement regroupées sous les formes *goût(pas_suffisant)*, *alcoolisé(faible)* et *amer(trop).*

4.2.3. Exemple et effet de codage

L'ensemble des étapes décrites dans les sections précédentes est automatisé grâce à des tables de reformatage construites dans le logiciel INFOTRANS. Le Tableau 26 montre le résultat du codage à partir d'un extrait de commentaires collectés auprès des consommateurs.
Nous pouvons observer la transformation de la phrase en unités minimales codées qui seront davantage adaptées au traitement statistique.

Tableau 26 : Exemple de codage

Commentaires collectés	Commentaires codés
un peu trop claire trop fade, la mousse n'est pas onctueuse	clair(trop) fade(trop). mousse onctueux(faible)
est agréable, a du parfum, je la trouve légèrement sucrée	agréable. aromatisé. sucré(faible)
bière sans qualités particulières, ne sort pas de l'ordinaire	qualité(faible). classique(faible)
goût peu prononcé, pas assez mousseuse, je lui trouve un goût trop sucré	goût(faible). mousse(pas_suffisant). sucré(trop)

Le codage est un moyen de diminuer la dispersion du vocabulaire tout en essayant de perdre le moins possible d'information. Mais qu'en est-il vraiment de son effet sur la distribution des fréquences des formes ?

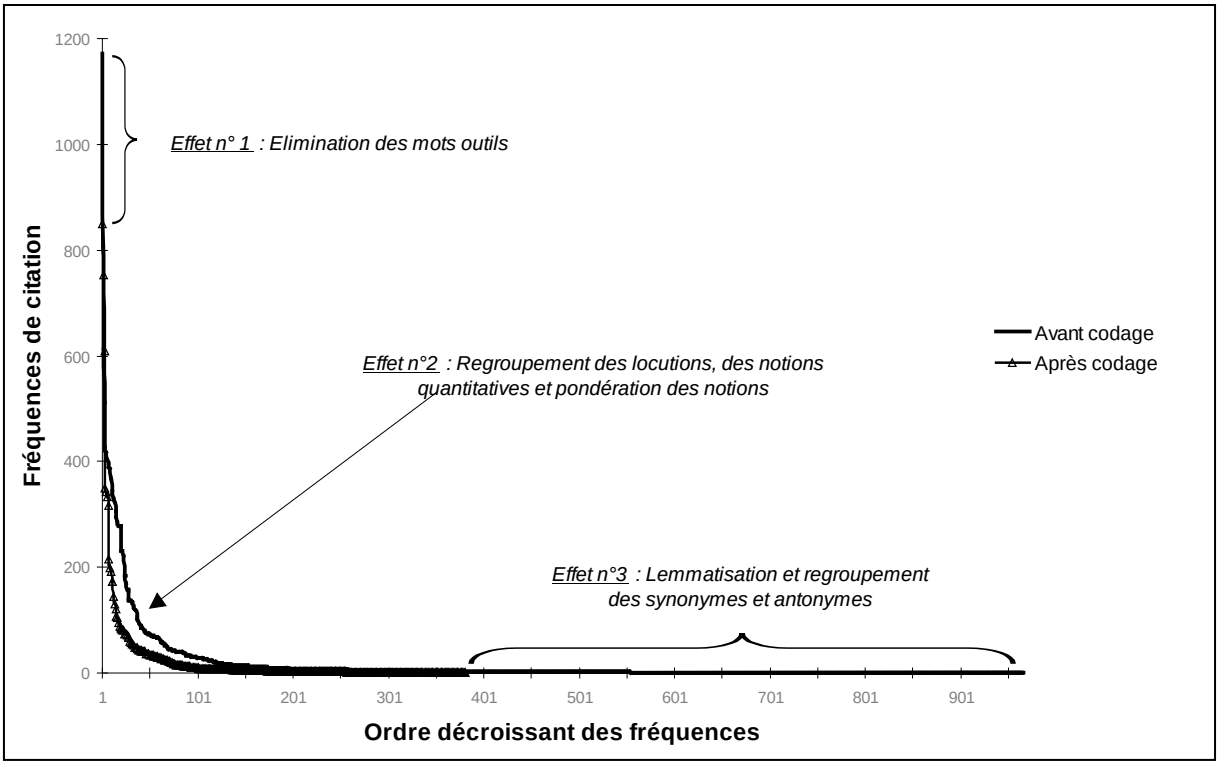


Figure 23 : Effet du codage

La Figure 23 montre que le codage change globalement la distribution du vocabulaire. En effet, nous pouvons observer de façon très nette trois effets principaux :

- effet n°1 : la disparition des mots outils (Zone A de la Figure 6),
- effet n°2 : la courbe après codage est au-dessous de la courbe avant codage en ce qui concerne l'information intéressante (Zone II de la Figure 6). Nous avons attribué cet effet à la normalisation du vocabulaire (regroupements de locutions, des notions quantitatives et pondération des notions). Dans le cas concret de la forme goût, nous observons une diminution de 323 occurrences en raison d'un regroupement sous les formes arrière-goût mais aussi goût(fort), goût(faible), …
- effet n°3 : la disparition des différentes formes fléchies, des synonymes et antonymes (Zone III de la Figure 6).

4.3. Traitement statistique et représentation graphique

Parmi les différents choix de traitements statistiques, nous avons choisi l'analyse des cooccurrences (voir section 2.1.3).
Deux concepts nous intéressent à travers le traitement statistique et la représentation graphique :

- l'aspect base de données : sauvegarde du patrimoine (assurer une rentabilité des données)
- l'aspect construction graphique : utilisation ponctuelle des données, mise à profit rapide.

C'est en partie pour cette raison que nous avons décidé qu'il n'était pas rentable d'utiliser DATAVIEW uniquement pour les calculs de cooccurrence. Ces derniers peuvent être facilement réalisés par l'intermédiaire d'un gestionnaire de base de données approprié. Pour cela, nous en avons évalué plusieurs et nous avons choisi le SGBDR[32] TEXTO™ de CHEMDATA car il offre non seulement la possibilité de stocker les commentaires sous la forme d'une base de données relationnelle mais il crée également des index sur l'ensemble des occurrences et des cooccurrences **[CHEM96]**.
Il s'agit des index spéciaux qui sont réalisés sur les commentaires codés. Ils fournissent la liste de l'ensemble des paires de termes avec leur fréquence d'apparition. Nous obtenons deux types d'information : la liste des formes contenant tous les mots avec leur fréquence d'apparition et la liste des paires contenant toutes les associations de deux mots avec leurs fréquences d'apparition (voir ANNEXE 6).

Il ne suffit pas d'avoir des données et le calcul statistique pour les valoriser par une prise de décision. En effet, les données brutes ne restituent pas toute l'information. Les relations construites par l'ensemble des données sont davantage riches. Autrement dit, pour J. BERTIN *l'information utile à la décision est faite des relations d'ensemble* **[BERT77]**.
Notre représentation graphique devra donc nous aider à découvrir ces relations d'ensemble. C'est dans cet esprit qu'une application de construction automatique de graphe a été développée. Plusieurs éléments ont guidé notre recherche en fonction de la spécificité des données et de nos attentes en matière de représentation graphique (voir section 3.3), à savoir :

- représentation des mots descriptifs avec leur pondération
- représentation sous forme de réseau de connexion
- comparaison des représentations

[32] Système de Gestion de Bases de Données Relationnelles

Différentes solutions que nous n'exposerons pas ici ont été envisagées à partir des graphes de CANDIDE et de MATRISME en combinant toutes les possibilités pour arriver à regrouper les trois attentes citées ci-dessus. La représentation en réseau est un bon moyen de montrer les structures des commentaires détruites lors de la segmentation (voir section 2.1.3). Elle ne l'est plus dès que le nombre des éléments augmente car la figure devient rapidement complexe, illisible et intransformable (voir section 2.2.4). Il est donc indispensable de passer par une phase de simplification du réseau. E. BOUTIN y parvient en effectuant des seuils sur les fréquences des cooccurrences et sur le nombre de liaisons par élément constitutif du réseau **[BOUT96]**. Notre approche consiste à représenter des morceaux de réseau autour d'un élément principal (forme pôle). La représentation, appelée graphe des mots étoilés est ainsi plus simple à lire et à interpréter (voir Figure 28).

Le programme DANOTEX[33] va nous permettre de construire automatiquement les graphes en :

- mettant en évidence les termes les plus fréquents
- mettant en évidence les liaisons les plus fortes entre deux termes
- utilisant la représentation en réseau de mots étoilés
- obtenant une représentation synthétique
- déduisant une interprétation rapide et simple

Il se lance à partir d'un fichier EXCEL. Ce dernier ouvre un classeur qui comporte à l'ouverture, une feuille *Présentation*, une feuille *Données* et une feuille *Réf P*. Ces trois feuilles font partie de la structure de base du programme. En voici leur désignation :

Tableau 27 : Description des feuilles contenues dans le fichier EXCEL de départ

Nom de la feuille	Désignation
Présentation	Ecran de présentation avec menu principal
Données	Tableaux dans lesquels seront inscrites les données au cours du calcul.
Réf P	Ensemble des fichiers produits associés aux noms de produits et aux nombres de commentaires

[33] Développé pour le TEPRAL en Visual Basic pour EXCEL de façon à pouvoir faire intervenir d'autres applications étrangères à EXCEL. En effet, la particularité de Visual Basic pour Applications est que non seulement il peut se servir des ressources de Windows mais il peut aussi se servir des ressources des autres applications Windows qui sont conçues pour reconnaître le langage de Microsoft Visual Basic pour applications.

Le menu principal se présente de la façon suivante :

Fichier	Editions	Graphe	Comparaison
↳ Importer pour un produit ↳ Modifier ↳ Graphe suivant... ↳ Aperçu avant impression ↳ Imprimer ↳ Fermer	↳ Codes sources ↳ Destruction d'une feuille ↳ Destruction du graphe et des tableaux	↳ Combinaison mots ↳ Combinaison produits	↳ Comparer

Figure 24 : Synoptique des menus de DANOTEX

Trois types de fonctions sont principalement distingués :

- la combinaison des mots
- la combinaison des produits
- la comparaison des graphes

4.3.1. Combinaison des mots

Cette étape est déclenchée dans le but d'analyser le vocabulaire pour un produit donné, un test donné et une question donnée.

Paramétrage du graphe

Le graphe des mots combinés se construit à partir d'un fichier d'index importé depuis le logiciel TEXTO™. Ce fichier appelé fichier PRN[34] correspond à un test et une question précise sur un produit.

[34] Sigle de l'extension sous Windows

Le programme nous offre ensuite la possibilité de choisir une des combinaisons ou classes de mots déjà mémorisées dans des tables paramétrées (liste des cinq mots les plus fréquents parmi l'ensemble des classes, liste des cinq mots les plus fréquents en dehors des mots choisis dans les classes) :

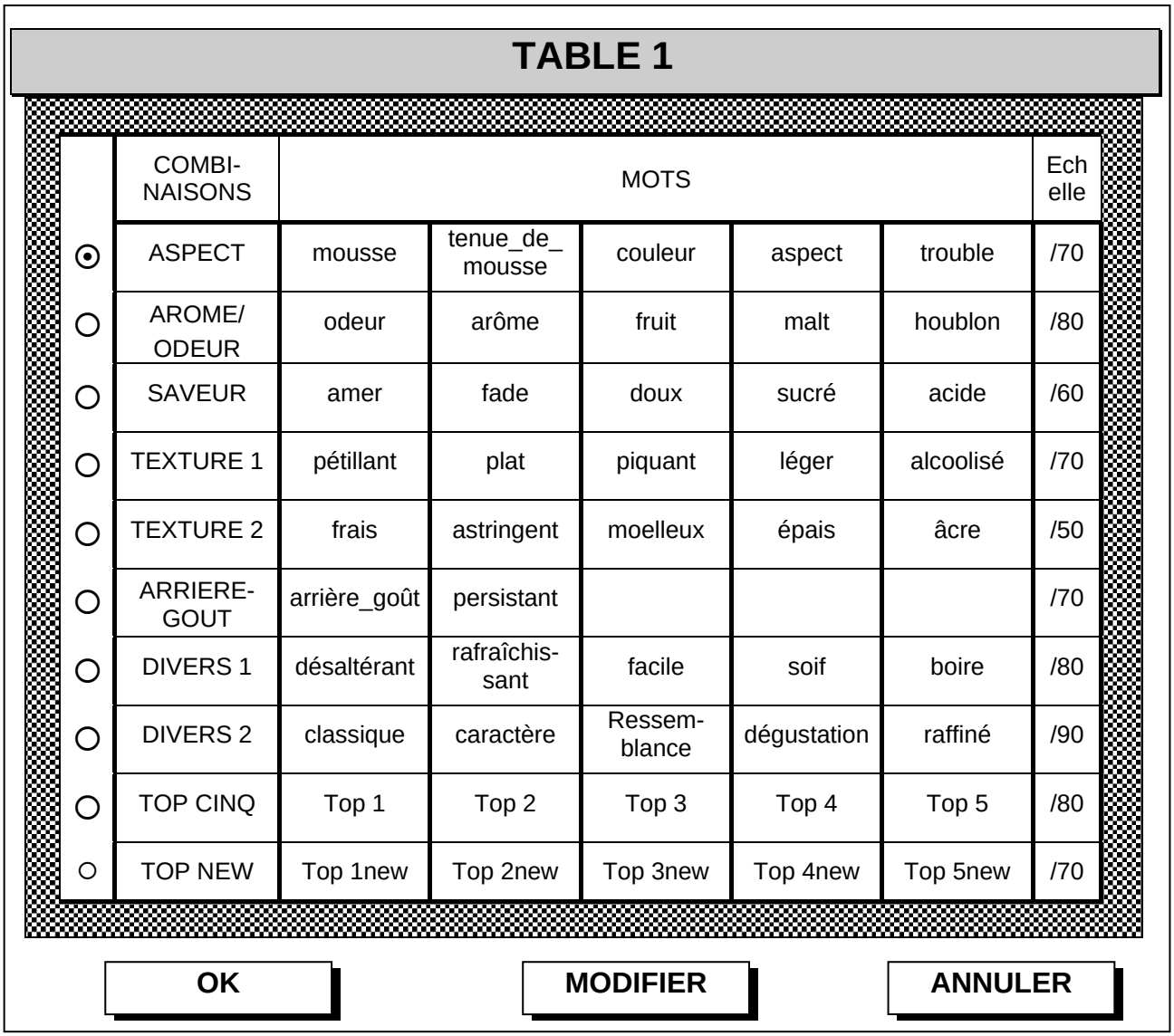


Figure 25 : Choix des combinaisons de mots

Chaque table peut être éditée de façon à modifier les combinaisons. Si l'utilisateur désire garder une table sans la modifier, il a la possibilité d'en créer une nouvelle.

Pour chaque combinaison, nous devons fixer un maximum d'échelle. En effet, la répartition des mots sur le graphe est réalisée en fonction de leurs pourcentages de citation. Il s'agit donc ici de réaliser un ajustement de l'échelle.

Le choix d'une combinaison de mots entraîne la préparation de la construction du graphe. L'utilisateur doit sélectionner une série de paramètres dans la boîte de dialogue suivante :

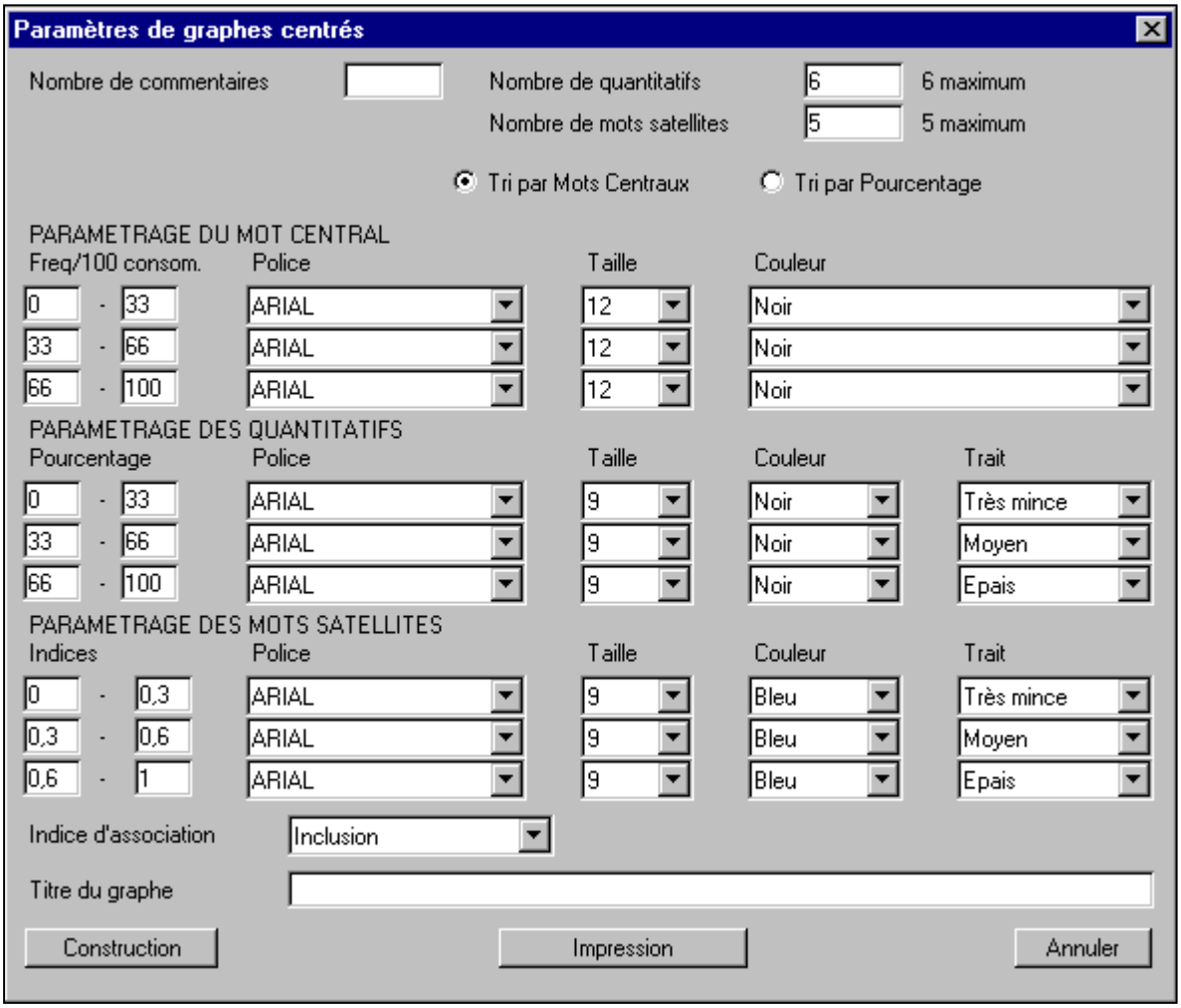


Figure 26 : Paramètres des combinaisons de mots

En premier lieu, l'utilisateur doit inscrire le nombre de commentaires qui composent le fichier de données. Ce nombre est important pour les calculs de fréquences et d'indice d'association (voir section 2.1.3).
Afin de faciliter la lecture du graphe, le positionnement des mots centraux peut se faire selon deux modes :

- disposition sur l'axe vertical par ordre alphabétique
- disposition sur l'axe vertical par ordre croissant du pourcentage

Pour la deuxième, il s'agit d'un tri numérique sur les pourcentages (dans la feuille de données) effectué avant de lancer la construction du graphe.
D'autre part, cette disposition introduit deux traits en pointillés verticaux qui partagent le graphe en trois zones. Ceci ne sera pas appliqué dans le cas de la disposition par ordre alphabétique. Cette séparation permet de mieux distinguer les groupes d'éléments faiblement cités, moyennement cités et fortement cités.

Leur position est déterminée à partir du calcul suivant :

1) moyenne des pourcentages de citation = M (sur l'axe vertical)
2) borne supérieure = M + M/2
3) borne inférieur = M - M/2
4) le premier trait se positionnera entre le mot central qui a un pourcentage de citation inférieur à la borne supérieure et le mot central qui a un pourcentage de citation supérieur à la borne inférieure
5) le deuxième trait se positionnera entre le mot central qui a un pourcentage de citation inférieur à la borne supérieure et le mot central qui a un pourcentage de citation supérieur à la borne inférieure

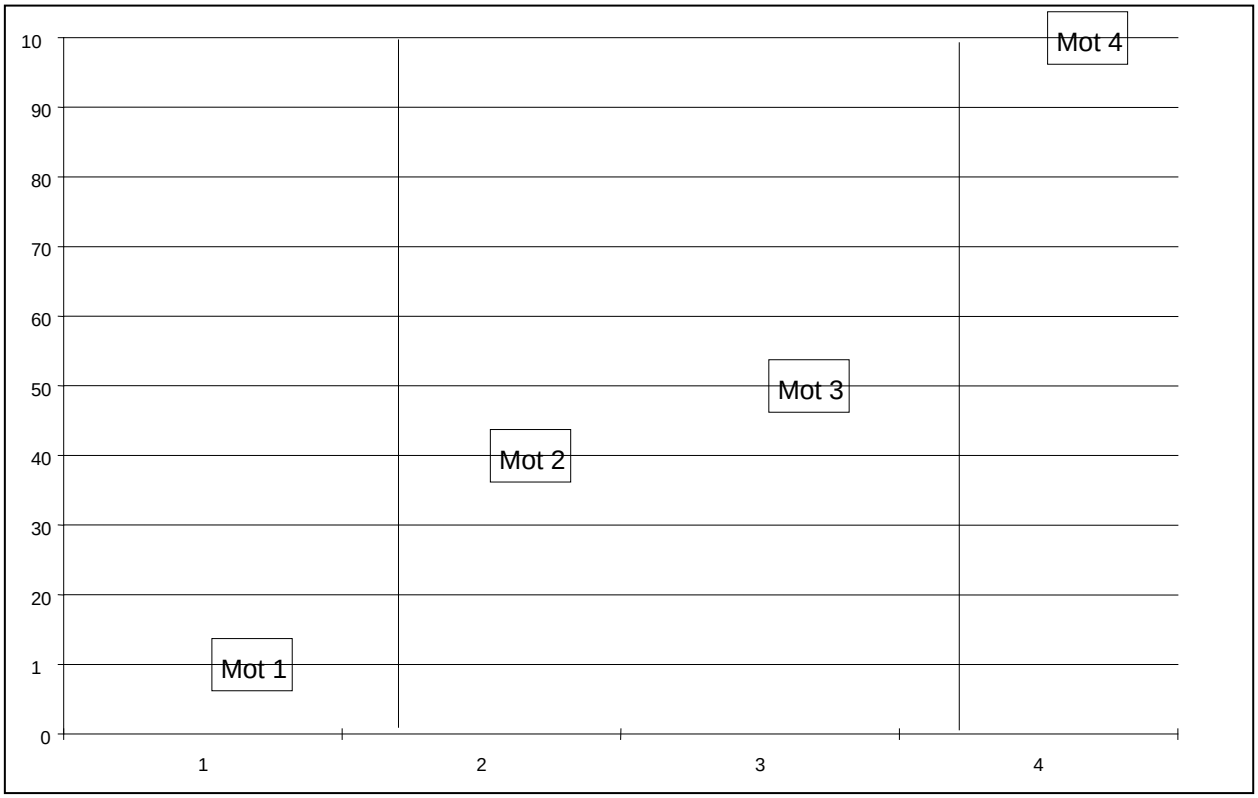


Figure 27 : Disposition des mots centraux sur le graphe

REMARQUE :
La deuxième solution sera préférée si nous souhaitons effectuer une comparaison de graphe.

Ensuite, il peut paramétrer le nombre de qualificatifs et de mots satellites, la police de caractère la taille et la couleur des mots ainsi que l'épaisseur des traits.
Il peut également choisir l'indice d'association (Inclusion, Jaccard, Corrélation, Russel & Rao), fixer l'échelle verticale et donner un titre au graphe.

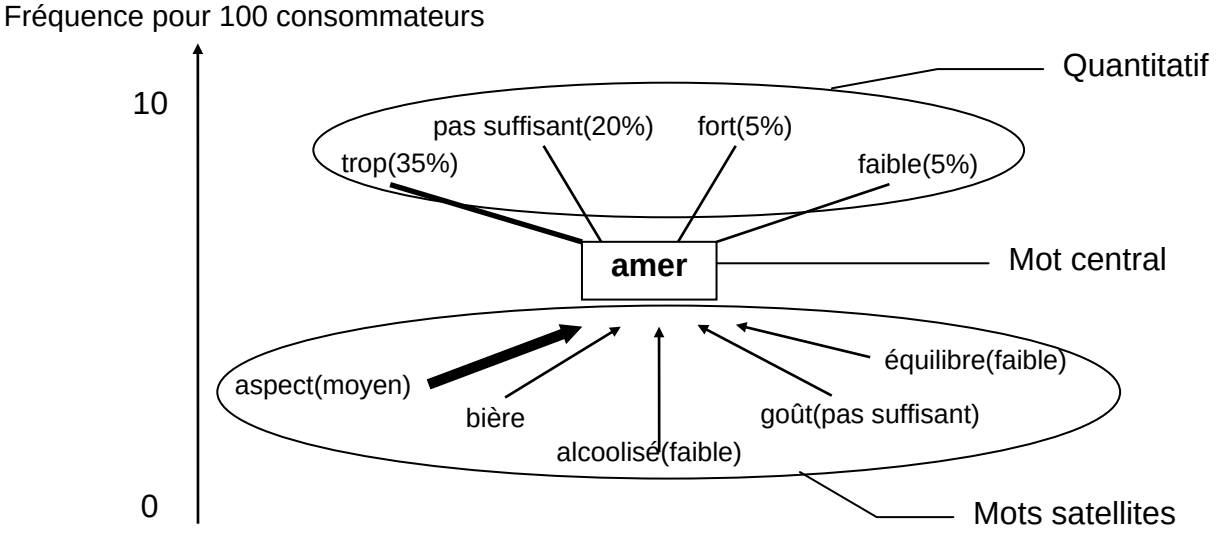
Le paramétrage distingue trois types de mots :

⇒ le **mot central** qui se situe comme son nom l'indique au centre, en gras et encadré (voir exemple ci-dessous).

⇒ les **quantitatifs** qui se situent au-dessus du mot central. Il y en a six au maximum. Ils sont exprimés en pourcentage d'association (voir tableaux plus loin). Ils représentent le contexte de citation du mot central. Dans l'exemple ci-dessous, l'amertume est citée comme insuffisante dans 20% des cas.

⇒ les **mots satellites** qui se situent au-dessous du mot central. Il y en a cinq au maximum. Ils sont déterminés par un calcul de l'indice d'association paramétré (voir tableaux plus loin). Ils représentent les notions les plus associées au mot central parmi les commentaires. Dans l'exemple ci-dessous, l'amertume est entre autre associée dans ce commentaire à l'aspect du produit.

EXEMPLE :



REMARQUE :
Le Tableau 28 rappelle les différents types d'information mis en évidence par les quatre indices disponibles dans DANOTEX (voir section 2.1.3) :

Tableau 28 : Indices d'association utilisés dans DANOTEX

Indices	Information mise en évidence
Indice de Jaccard	Il favorise l'apparition des paires présentant une forte intensité de lien avec des fréquences de co-apparition relativement élevées.
Indice de Russel & Rao	Il met en évidence la fréquence relative au nombre de commentaires de la paire.
Coefficient de corrélation	Il met en évidence deux types d'information : en positif les mots qui apparaissent toujours ensembles et en négatif les mots qui n'apparaissent jamais.
Inclusion	Il favorise l'apparition des paires présentant une forte intensité de lien avec des fréquences de co-apparition relativement faibles. Il donne un sens à la paire (montre si le mot satellite est le plus souvent cité avec le mot central ou pas)

Calcul du graphe

La construction automatique est transparente pour l'utilisateur. Elle comporte plusieurs étapes que nous allons détailler.

Première étape

Elle consiste à rechercher les informations suivantes dans le fichier de paramétrage du graphe (macro-commande EXCEL) et le fichier d'import des formes et paires de TEXTO™ (*.PRN) :

- nombre de consommateurs qui correspond au nombre de commentaires (voir paramètres du graphe)
- repérage des mots centraux choisis pour établir la liste des quantitatifs associés ainsi que les fréquences de ces associations (voir ANNEXE 6).

Deuxième étape

Elle réalise le calcul du sous-total des fréquences du mot central sur l'axe vertical ainsi que le calcul de la position du mot central sur l'axe vertical (voir Table 1).

Table 1

Liste de mots retrouvés dans le fichier TEXTO™, parmi les

Mots	Fréquences	Fréquence/nb de consommateurs
acide	3	(4/40) x 100 = **10**
acide(moyen)	1	
SOUS TOTAL **acide**	4	
amer	7	(20/40) x 100 = **50**
amer(faible)	1	
amer(fort)	1	
amer(pas suffisant)	4	
amer(trop)	7	
SOUS TOTAL **amer**	20	
fade	2	(3/40) x 100 = **7,5**
fade(moyen)	1	
SOUS TOTAL **fade**	3	
sucré(moyen)	1	(3/40) x 100 = **7,5**
sucré(pas suffisant)	1	
sucré(trop)	1	
SOUS TOTAL **sucré**	3	
Nombre de consommateurs	40	

La Table 1 est obtenue à partir d'une extraction du fichier d'édition des paires par ordre alphabétique de TEXTO™ (voir exemple en ANNEXE 6).

Troisième étape

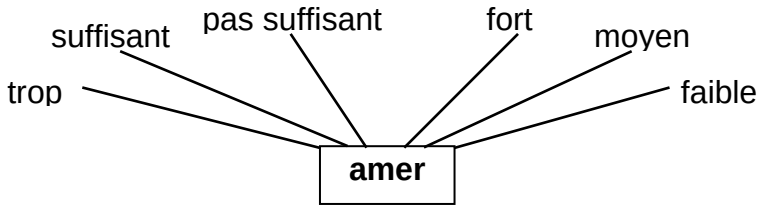
Elle calcule les pourcentages des quantitatifs (voir Table 2)

Table 2

Quantitatifs	Pourcentages
acide(moyen)	(1/4) x 100 = **25**
amer(faible)	(1/20) x 100 = **5**
amer(fort)	(1/20) x 100 = **5**
amer(pas suffisant)	(4/20) x 100 = **20**
amer(trop)	(7/20) x 100 = **35**
fade(moyen)	(1/3) x 100 = **33**
sucré(moyen)	(1/3) x 100 = **33**
sucré(pas suffisant)	(1/3) x 100 = **33**
sucré(trop)	(1/3) x 100 = **33**

Quatrième étape

Elle dispose les quantitatifs au-dessus du mot central.



Cinquième étape

Elle recherche les informations dans le fichier des formes et des paires de TEXTO™ (voir fichier *.PRN en ANNEXE 6). Il s'agit de repérer les formes associées (mots satellites) au mot central et à ses nuances quantitatives avec leurs fréquences respectives (voir Table 3).

Table 3

Paires brutes	Fréquences brutes
acide- pétillant	1
acide-équilibré(trop)	2
acide(moyen)-pétillant	1
acide(moyen)-désaltérant	1
amer-alcoolisé(pas suffisant)	2
amer-aspect(moyen)	1
amer-bière	1
amer-équilibré	3
amer(faible)-pétillant	1
amer(fort)-aspect(moyen)	1
amer(pas suffisant)-bière	1
amer(pas suffisant)-pétillant	1
amer(pas suffisant)-couleur	1
amer(trop)-couleur	4
amer(trop)-pétillant	1
amer(trop)-équilibre	2
fade-caractère	1
fade-goût(faible)	1
fade-frais(pas suffisant)	1
fade(moyen)-caractère	1
fade(moyen)-goût(faible)	1
sucré(moyen)-équilibré	1
sucré(moyen)-raffiné(moyen)	1
sucré(pas suffisant)-raffiné(moyen)	1
sucré(trop)-boire(faible)	1

La Table 3 (paires et fréquences brutes) est une extraction du fichier d'édition des paires par ordre alphabétique de TEXTO™ (voir exemple en ANNEXE 6).

Sixième étape

Elle calcule les fréquences d'association entre le mot central sans nuance quantitative avec les mots satellites (voir Table 4)

Table 4

Paires nettes	Fréquences nettes
acide-pétillant	1+1 = **2**
acide-équilibré(trop)	**2**
acide-désaltérant	**1**
amer-alcoolisé(pas suffisant)	**2**
amer-aspect(moyen)	1+1 = **2**
amer-bière	1+1 = **2**
amer-équilibré	3+2 = **5**
amer-pétillant	1+1+1 = **3**
amer-couleur	4+1 = **5**
fade-caractère	1+1 = **2**
fade-goût(faible)	1+1 = **2**
fade-frais(pas suffisant)	**1**
sucré-boire(faible)	**1**
sucré-équilibré	**1**
sucré-raffiné(moyen)	1+1 = **2**

La Table 4 (paires et fréquences nettes) est obtenue à partir de la table 3 : nous additionnons les paires comportant les associations entre les différentes nuances quantitatives de mots centraux et des mots satellites identiques.

Septième étape

Elle prépare des informations pour les calculs d'indice en repérant les mots satellites et de leurs fréquences (voir Table 5)

Table 5

Mots satellites	Fréquences
alcoolisé(pas_suffisant)	2
aspect(moyen)	2
bière	2
boire(faible)	1
caractère	2
couleur	5
désaltérant	11
frais(pas_suffisant)	1
goût(faible)	2
pétillant	5
raffiné(moyen)	2
équilibré	6
équilibré(trop)	2

La Table 5 est une extraction du fichier d'édition des paires par ordre alphabétique de TEXTO™ (voir exemple en ANNEXE 6).

Huitième étape

Elle calcule les valeurs d'indice (voir Table 6).

Table 6

Paires	(A)	(B)	(C)	(D)	R[35]	J[36]	C[37]	I[38]
Acide Nb de mots satellites représentés/nb de mots satellites totaux = 3/13 Information représentée : (3/13) x 100 = 23 %								
acide-pétillant	2	4-2=**2**	5-2=**3**	40-7=**33**	0.05	0.285	0.377	0.5
acide-équilibré(trop)	2	4-2=**2**	2-2=**0**	40-4=**36**	0.05	0.5	0.688	1
acide-désaltérant	1	4-1=**3**	11-1=**10**	40-4=**26**	0.025	0.25	-0.01	0.25
Amer Nb de mots satellites représentés/nb de mots satellites totaux = 5/13 Information représentée : (5/22) x 100 = 38 %								
amer-alcoolisé(pas suffisant)	2	20-2=**18**	2-2=**0**	40-20=**20**	0.05	0.1	0.229	1
amer-aspect(moyen)	2	20-2=**18**	2-2=**0**	40-20=**20**	0.05	0.1	0.229	1
amer-bière	2	20-2=**18**	2-2=**0**	40-20=**20**	0.05	0.1	0.229	1
amer-équilibré	5	20-5=**15**	6-5=**1**	40-21=**19**	0.125	0.238	0.280	0.83
amer-pétillant	3	20-3=**17**	5-3=**2**	40-22=**18**	0.075	0.136	0.075	0.6
amer-couleur	5	20-5=**15**	5-5=**0**	40-20=**20**	0.125	0.25	0.377	1
Fade Nb de mots satellites représentés/nb de mots satellites totaux = 3/13 Information représentée : (5/22) x 100 = 23 %								
fade-caractère	2	3-2=**1**	2-2=**0**	40-4=**36**	0.05	0.666	0.805	1
fade-goût(faible)	2	3-2=**1**	2-2=**0**	40-3=**37**	0.05	0.666	0.805	1
fade-frais(pas suffisant)	1	3-1=**2**	1-1=**0**	40-5=**35**	0.025	0.333	0.561	1
Sucré Nb de mots satellites représentés/nb de mots satellites totaux = 3/13 Information représentée : (5/22) x 100 = 23 %								
sucré-boire(faible)	1	3-1=**2**	1-1=**0**	40-3=**37**	0.025	0.333	0.562	1
sucré-équilibré	1	3-1=**2**	6-1=**5**	40-8=**32**	0.025	0.125	0.146	0.33
sucré-raffiné(moyen)	2	3-2=**1**	2-2=**0**	40-4=**36**	0.05	0.666	0.805	1

[35] Russel & Rao (voir formule 2)
[36] Jaccard (voir formule 4)
[37] Corrélation (voir formule 5)
[38] Inclusion (voir formule 6)

Avec :

(A) = fréquence de la paire (mot central / mot satellite)

(B) = fréquence mot central - fréquence de la paire ou (A)

(C) = fréquence mot satellite - fréquence de la paire ou (A)

(D) = nombre de consommateurs - A - B - C

Neuvième étape

Elle trie et conserve des paires (cinq maximum) qui ont les plus fortes valeurs d'indice. L'épaisseur du trait varie en fonction de la valeur de l'indice et des paramètres qui ont été fixés (voir Figure 26).

Elle donne un sens à l'association dans le cas de l'inclusion. Pour l'inclusion, le sens du lien entre le mot central et le mot satellite dépend de leurs fréquences respectives :

- si fréquence(mot central) > fréquence(mot satellite), alors la flèche va dans le sens mot satellite $\Rightarrow$ mot central,
- Si fréquence(mot central) < fréquence(mot satellite), alors la flèche va dans le sens mot central $\Rightarrow$ mot satellite.

Pour les autres indices, la liaison n'est pas orientée. Elle est seulement représentée par un trait.

Dixième étape

Elle calcule le pourcentage d'information représenté pour les mots satellites (voir Table 6) :

A partir de la table 6, le nombre de lignes correspond au nombre de mots satellites.
A partir de la feuille de données dans la zone correspondant aux mots satellites ou à partir de la table 5 on connaît le nombre de mots maximum qui seront représentés.

Le pourcentage d'information représenté pour les mots satellites est alors calculé de la façon suivante :

$$\frac{\text{Nombre de mots satellites maximal fixé par l'utilisateur}}{\text{Nombre de mots satellites existants (pour un fichier d'index donné)}} \text{ X } 100$$

Lorsque le nombre de mots satellites existants est inférieur au nombre de mots satellites maximal fixé par l'utilisateur, ce dernier prendra la valeur du premier.

Les différents éléments du graphe peuvent être enfin disposés sur une feuille de graphe EXCEL.

Le fichier de résultat comporte plusieurs feuilles en plus des trois feuilles du départ et de la feuille de données PRN. Ces dernières ont participé à l'élaboration du graphe. En voici leur désignation :

Nom de la feuille	Contenu
Données PRN	Fichier des formes et paires de TEXTO™
Table	Combinaisons des produits
Tableau 7	Mots satellites choisis et valeurs d'indice
Tableau 6	Valeurs des indices
Tableau 5	Fréquences des mots satellites
Tableau 4	Fréquences nettes des paires
Tableau 3	Fréquences brutes des paires
Tableau 2	Pourcentages des quantitatifs
Tableau 1	Fréquences des mots centraux et fréquences des mots centraux pour 100 consommateurs
Graphe	Représentation graphique des mots étoilés
Données	Synthèse des données utiles à la construction du graphe

L'ensemble de ces feuilles est stocké dans le même classeur EXCEL. Chacune d'entre-elles peut être sélectionnée et détruite.

Lorsque la construction d'un graphe est validée, les feuilles *Données PRN, Table* et de *Tableau 1* à *Tableau 7* sont détruites. Ces dernières représentent en effet des fichiers temporaires de calcul.

Représentation et Interprétation du graphe

Le programme construit donc le graphe des mots étoilés à partir du fichier d'index importé et des critères que nous avons paramétrés.
Le graphique est stocké dans le classeur EXCEL sur la feuille *Graphe* avec la feuille *Données* qui a permis de le créer.

Il se présente de la façon suivante :

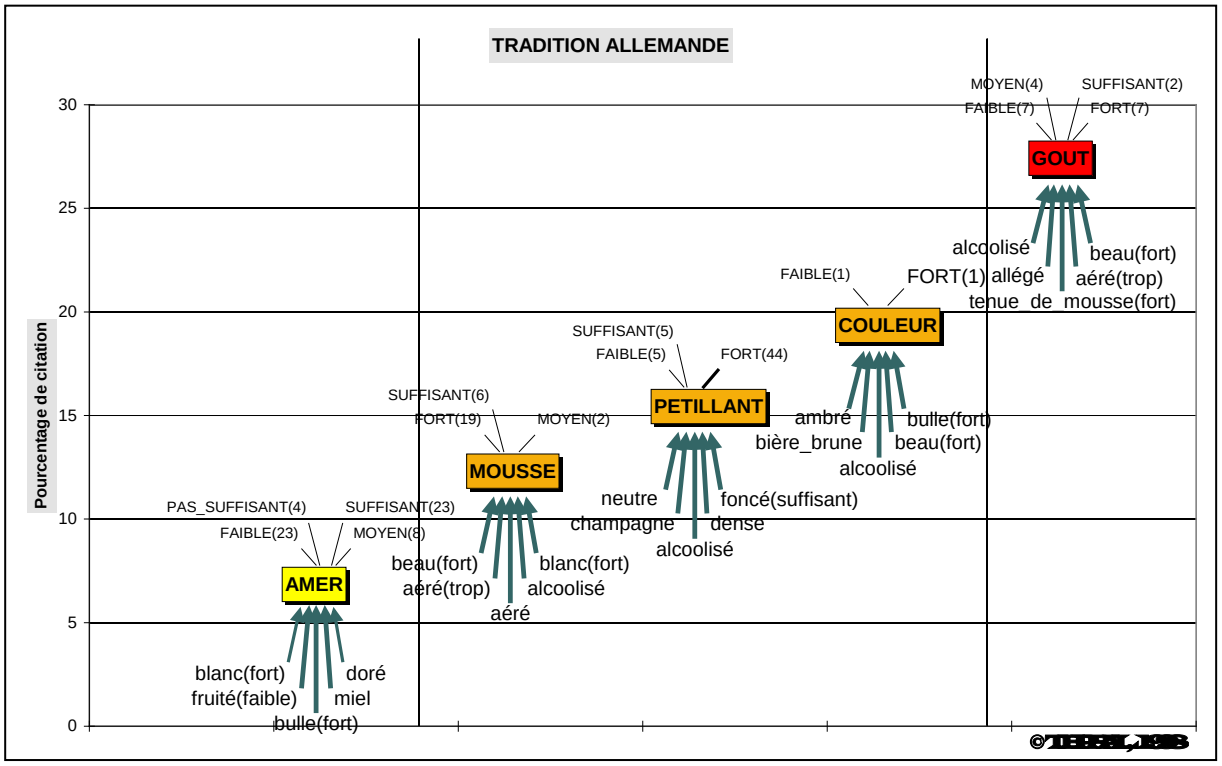


Figure 28 : Graphe des mots étoilés (option combinaison des mots)

Cinq mots centraux au maximum sont disposés par ordre de pourcentage de citation croissant (ou ordre alphabétique) suivant l'axe horizontal et en fonction de leur fréquence pour 100 consommateurs suivant l'axe vertical.
Dans les cas où le mot central n'est pas retrouvé dans le fichier PRN, il est considéré comme nul. Il est donc placé sur la ligne correspondant à zéro.

Les graphes des mots caractéristiques mettent en évidence trois types d'informations complémentaires :

⇒ Mots centraux : ce sont les termes les plus fortement cités de façon spontanée pour un produit choisi : ils sont en encadré avec une trame jaune ou gris très clair (pour les moins cités), orange ou gris (pour les moyennement cités) et rouge ou gris foncé (pour les plus cités). Ils peuvent être répartis en trois groupes grâce aux séparations verticales (voir Figure 27).

Par exemple : la Figure 28 montre les cinq mots les plus cités (*amer*, *mousse*, *pétillant*, *couleur* et *goût*), pour la bière *Tradition Allemande* dans la question sur les qualités. Ils sont disposés dans l'ordre croissant des fréquences de citation.

⇒ Mots quantitatifs (au-dessus du mot central) : ils précisent la nuance de citation avec laquelle le mot choisi a été cité. Ils sont exprimés en pourcentage du mot choisi. Ce sont les termes qui ont été regroupé dans deux types d'échelles (voir section 4.2.2).

L'épaisseur du trait varie en fonction de la valeur de ce pourcentage.

Par exemple : la Figure 28 montre que le mot *pétillant* est cité à 44 % en tant que fortement pétillant.

⇒ Mots satellites (au-dessous du mot central) : ils précisent le contexte de citation dans lequel le mot choisi a été cité. Les cinq mots les plus liés (calculé à partir d'un indice de cooccurrence) au mot choisi apparaissent.

L'épaisseur de la flèche indique le degré d'intensité de la liaison : une flèche épaisse signifie que les deux mots sont cités ensembles dans la majorité des commentaires, une flèche d'épaisseur moyenne signifie que les mots sont cités aussi bien ensembles que seuls et enfin, une flèche fine signifie que les deux mots sont peu souvent co-cités.

D'autre part, si le calcul de cooccurrence a été fait à partir de l'indice d'inclusion, l'orientation de la flèche indique si le mot satellite est toujours associé au mot central ou non : la flèche sera dirigée dans le sens mot satellite, mot central dans le cas où l'occurrence du mot central serait supérieure à l'occurrence du mot satellite et inversement dans le cas où l'occurrence du mot central serait inférieure à l'occurrence du mot satellite.

Par exemple : la Figure 28 montre que le mot *couleur* est dans la majorité des commentaires co-cité avec les termes *ambré*, *bière brune*, *très bulleuse*, *très belle* et *alcoolisé*. La fréquence du mot couleur est dans tous les cas supérieure aux fréquences des mots satellites.

4.3.2. Combinaison des produits

Cette étape est déclenchée lorsque plusieurs produits pour un terme identique sont à comparer.

Paramétrage du graphe

Le graphe est construit à partir d'un même mot retrouvé dans plusieurs fichiers d'index que nous importons de TEXTO™.
Nous appelons produit la réponse à une question ouverte donnée, après la dégustation d'un produit donné, pour un test donné.

Le programme nous offre ensuite la possibilité de choisir une des combinaisons ou classes de produits déjà mémorisées dans des tables paramétrées :

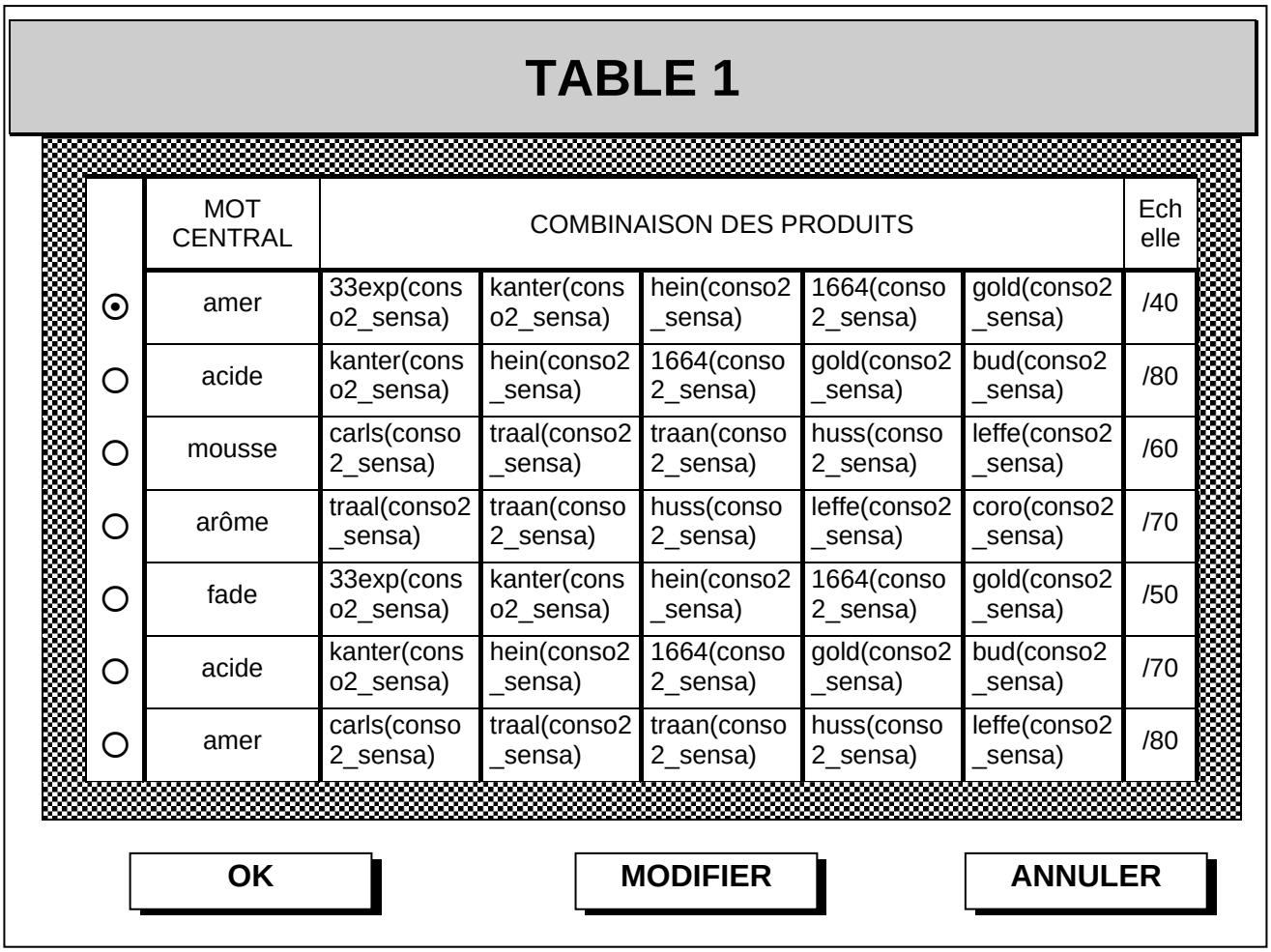


Figure 29 : Choix des combinaisons de produits

Chaque combinaison est inscrite dans une feuille de calcul nommée *Ref P* (dans le fichier EXCEL du programme) :

Tableau 29 : Liste des références produits

Nom combinaison	Chemin	Nombre commentaires	Nom explicite
1664(conso2_circo)	c:\l\special\danotex\circo16.prn	180	1664
1664(conso2_defau)	c:\l\special\danotex\defau16.prn	180	1664
1664(conso2_quali)	c:\l\special\danotex\quali16.prn	180	1664
1664(conso2_sensa)	c:\l\special\danotex\sensa16.prn	180	1664
33exp(conso2_circo)	c:\l\special\danotex\circo33.prn	179	33 export
33exp(conso2_defau)	c:\l\special\danotex\defau33.prn	179	33 export
33exp(conso2_quali)	c:\l\special\danotex\quali33.prn	179	33 export
33exp(conso2_sensa)	c:\l\special\danotex\sensa33.prn	179	33 export

Il est nécessaire d'enregistrer une combinaison de produits dans la feuille *Ref P* si c'est la première fois qu'elle est inscrite dans le programme. Cette dernière sera ensuite mémorisée.

Par convention, les codes des produits se déclinent de la façon suivante :

nom abrégé du produit(code du test_question)

EXEMPLE :

33exp(conso2_quali)
1664(conso2_defau)

Le choix d'une combinaison de produit entraîne la préparation de la construction du graphe. L'utilisateur doit sélectionner une série de paramètres dans la même boîte de dialogue que pour la combinaison des mots.

Calcul du graphe

La construction est basée sur le même principe que celle de la combinaison des mots à quelques exceptions près. Nous ne détaillerons donc pas les différentes étapes de construction.

Représentation et Interprétation du graphe

Le programme construit donc le graphe des mots étoilés à partir du fichier d'index importé et des critères que nous avons paramétrés.
Le graphique est stocké dans le classeur EXCEL sur la feuille *Graphe* avec la feuille *Données* qui a permis de le créer.
Il se présente de la façon suivante :

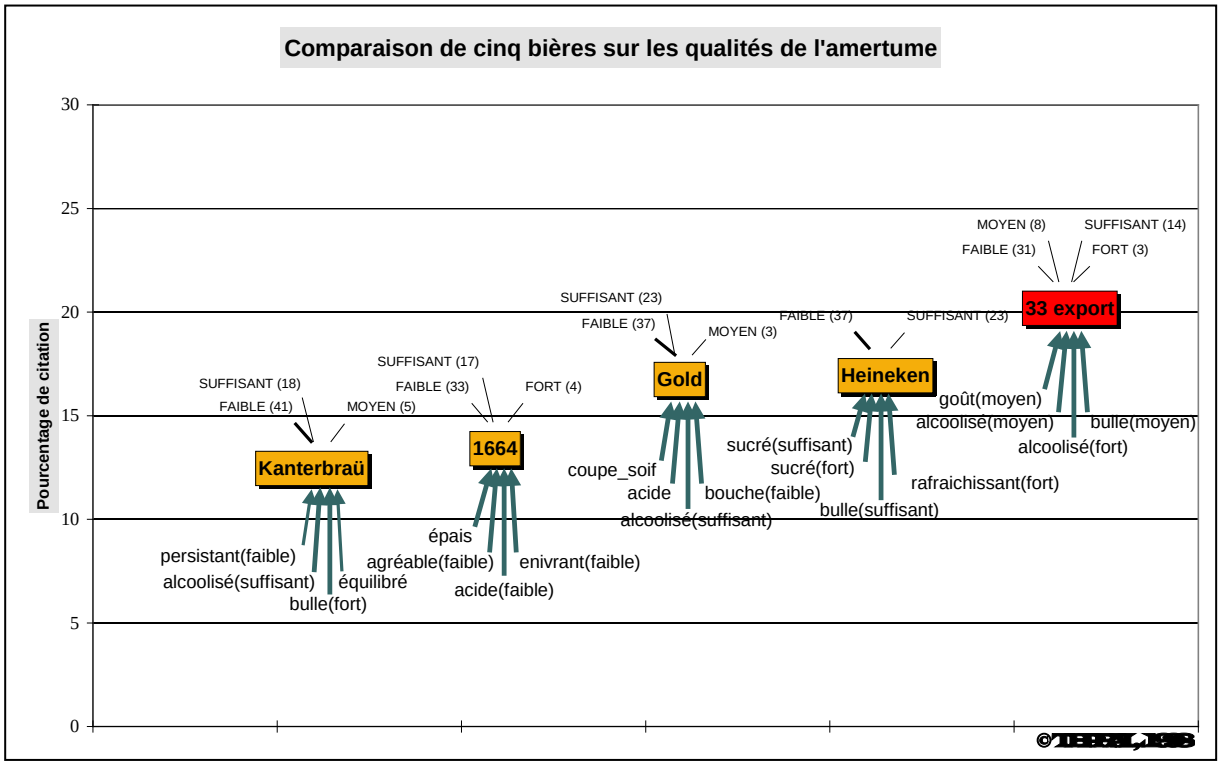


Figure 30 : Graphe des mots étoilés (option combinaison des produits)

Les règles d'interprétation vues au 4.3.1 permettent d'analyser l'information contenue sur ce graphe.
Sur la Figure 30, le mot central représente la citation du mot amertume et de ses flexions pour cinq bières différentes : Kanterbraü, 1664, Gold, Heineken et 33 export. Ces citations ont été extraites des commentaires libres sur les qualités de ces bières.
Les cinq mots centraux sont disposés par ordre alphabétique suivant l'axe horizontal et en fonction de leur fréquence pour 100 consommateurs suivant l'axe vertical.
Ici, les mots centraux ne sont par regroupés en trois classes puisque nous voyons qu'il n'y a que deux types de fréquences, les fortes (33 export) et les moyennes (Kanterbraü, 1664, Gold et Heineken).
D'une façon générale, l'amertume a été citée en tant que faible pour les cinq bières, de façon plus importante pour la Kanterbraü, Gold et la Heineken. Nous remarquons d'autre part que l'amertume a aussi été citée en tant que suffisant pour la Gold et la Heineken.
Sur ce graphe, nous avons choisi de limiter le nombre de mots satellites à quatre. Ils sont tous dans la majorité des commentaires co-cités avec le mot amertume. Enfin, les fréquences des mots centraux sont dans tous les cas supérieurs aux fréquences des mots satellites.

A travers la connaissance du consommateur, nous avions espéré au premier chapitre trouver des voies nouvelles dans la recherche d'innovation. Le deuxième chapitre nous a montré que le potentiel était là mais que la tâche ne serait pas simple pour le mener à bien.
Il est en effet, très difficile de recueillir de l'information auprès des consommateurs. L'étude du mode de questionnement n'a d'ailleurs été qu'en partie développée par manque de temps. Nous envisagerons les pistes de travail qui pourront être développées à ce sujet au cours chapitre III.
De plus, le traitement et l'analyse de cette information est une opération encore moins simple.

Grâce à une confrontation d'idées en provenance de divers horizons et de compétences pluridisciplinaires, nous sommes arrivés à mettre au point une méthode satisfaisante pour valoriser l'expression libre des consommateurs après dégustation de produit alimentaire.
La démarche adoptée nous semble tout à fait intéressante dans la mesure où nous sommes partis d'un problème très pratique sur la recherche de la connaissance du consommateur que nous avons essayé de mesurer grâce aux méthodes très théoriques des statistiques.
L'analyse des solutions existantes et proches de nos besoins nous a permis de mieux maîtriser les techniques et d'envisager d'autres solutions de traitement pour nos données. Nous avons trouvé un grand nombre d'idées nouvelles grâce à l'association de plusieurs domaines tels que l'analyse sensorielle, la lexicométrie, la veille technologique, la bibliométrie, …
Ce mélange de compétences a en définitive enrichi considérablement notre travail et a ouvert de nombreuses autres perspectives de recherche.

D'une manière globale, cette méthode s'inscrit dans une démarche classique de traitement de l'information telle qu'on peut l'envisager notamment en veille technologique. Nous retrouvons en effet les différentes étapes de collecte, traitement, d'analyse et validation, de diffusion et de capitalisation (voir Tableau 30). Nous allons d'ailleurs voir dans le chapitre III, comment les étapes de diffusion et de capitalisation pourrons être envisagées dans le cas des commentaires libres de consommateurs.
Ce parallèle permet également de montrer que notre démarche s'inscrit dans un processus d'intelligence économique. Or, la gestion stratégique de l'information est devenue l'un des moteurs essentiels de la performance globale des entreprises et des nations. En effet, le processus de mondialisation des marchés contraint les agents économiques à s'adapter aux nouveaux équilibres qui s'établissent entre concurrence et coopération. Désormais, la conduite des stratégies industrielles repose largement sur la capacité des entreprises à accéder aux informations stratégiques pour mieux anticiper les marchés à venir et les stratégies des concurrents.
Dans ce sens, grâce à la valorisation des commentaires libres de consommateurs et la maîtrise de ce type d'information, nous sommes arrivés à mieux connaître leur perception des produits alimentaires. Sa mise en pratique, développée dans le chapitre III, a pu démontrer son utilité dans une démarche globale d'innovation en agroalimentaire.

Tableau 30 : Comparaison des stratégies de traitement

Etapes de la chaîne de traitement de l'information	Veille technologique	Commentaires libres
Collecte		
Sources	Observateurs	Consommateurs
Moyens	Equation de recherche	Questions ouvertes
Résultat	Information formelle et informelle	
Traitement		
Diminution du bruit	Reformatage	Codage
Outil	Statistiques	
Représentation graphique	Réseaux de cooccurrence	
Résultats	Indicateurs univariés et relationnels	
Analyse/Validation	Spécialistes du domaine + Spécialistes du traitement de l'information	
Diffusion		
Moyens	Groupware	
Cible	Décideurs	
Capitalisation	Base de connaissances, GED	

CHAPITRE III

CHAPITRE III : SYNTHESE ET CONCLUSION

Nous venons d'exposer en détail un projet de trois ans effectué dans le cadre d'une thèse en Sciences de l'Information et de la Communication au sein d'un grand groupe agroalimentaire français.
Nous souhaitons maintenant établir un bilan à la fois sur la réalisation des objectifs fixés (voir chapitre II, section 3), sur ce que nous avons apporté de nouveau pour la recherche et enfin sur les nouvelles voies ouvertes qu'il serait intéressant d'approfondir.

1. Réalisation des objectifs

D'une manière générale, la méthode que nous avons développée a répondu aux vœux de départ (voir chapitre II, section 3).
En effet, l'amélioration du codage et la simplification de la lecture des résultats ont complètement répondu aux attentes des industriels. De nombreux tests et applications de la méthode ont été conduits en collaboration avec plusieurs branches du groupe DANONE et avec le service du Développement des Brasseries Kronenbourg.

Un premier travail d'évaluation et de synthèse nous a permis d'émettre des grands axes de travail. Cette étape nous a fourni les bases pour développer une nouvelle méthode.

Optimiser le traitement

Les premières techniques (section 2. du chapitre II) de traitement des données textuelles proposaient des résultats dont la fiabilité n'était pas contrôlée[39].
Nous avons donc cherché avant tout à obtenir un traitement et un calcul simples et fiables à partir des lois statistiques sur les données textuelles (section 1.2. du chapitre II). Nous avons pu les trouver dans la littérature (sections 2.1.2. et 2.1.3. du chapitre II).

Optimiser la communication des résultats

De la même façon, nous sommes partis des représentations sous forme de réseau (sections 2.2.3. et 2.2.4. du chapitre II) qui étaient mieux appréhendés au niveau de l'interprétation que les analyses multidimensionnelles (section 2.2.1. du chapitre II). Nous les avons encore fait évoluer vers un mode de représentation simple à comprendre et à différents niveaux de lecture (section 4.3. du chapitre II). De cette manière, les résultats graphiques sont directement diffusables auprès des décideurs.

[39] L'interprétation était fonction de l'appréciation du lecteur.

Optimiser l'utilisation

La méthode que nous avons mise en place a été écrite. Les procédures permettent son application par un utilisateur novice de façon immédiate.

Adéquation avec les besoins

Les commentaires libres de consommateurs ont des caractéristiques spécifiques (section 1. du chapitre II) comme par exemple une répartition qui suit la loi de Zipf (section 1.2.1. du chapitre II). Grâce à une opération de codage (lemmatisation, élimination de mots vides, …) nous avons diminué la quantité de bruit au profit de l'information intéressante (section 4.2.1).

Les précédentes études sur les commentaires libres de consommateurs ont mis en évidence l'importance du contexte de citation (section 3.2.2.). Nous avons réussi à intégrer cette information aux résultats finaux (section 4.2.2.).

Rapidité

Enfin, concernant la rapidité, nous observons une différence entre les méthodes précédentes et celle que nous avons mise en place.
L'utilisation de SPAD.T par exemple, demandait peu de temps de traitement pour une interprétation et analyse extrêmement contraignante et longue.
Nous observons le phénomène inverse avec la méthode que nous avons présentée. Le traitement demande davantage de temps de travail alors que la phase d'interprétation et d'analyse est pratiquement immédiate. De plus, la répétition des études pour un même produit diminue le temps de traitement (capitalisation du vocabulaire).
EXEMPLE :

Temps de travail	SPAD. T	INFOTRANS/TEXTO/DANOTEX
Traitement	4 heures	8 heures, la première fois 4 heures, la deuxième fois
Interprétation	2 jours	4 heures

2. Application industrielle

Plusieurs études ont été menées au cours de ces trois années. Elles ont concerné plusieurs types de produits : bière, cidre, épicerie, confiserie.
Pour des raisons de confidentialité, les résultats ne seront pas détaillés. Il s'agit uniquement ici de mettre en évidence la fonctionnalité et l'efficacité de la méthodologie dans une optique industrielle.
Nous citerons trois types d'applications différentes pour lesquels la méthodologie a pu être mise en application.

2.1. Veille produit

La branche épicerie du groupe DANONE souhaitait cerner l'intérêt (appréciations hédoniques et intentions d'achat) des consommateurs français pour un produit américain nouveau sur le marché français. Un test consommateur a donc été conduit dans l'objectif de se faire une première idée de l'éventuel positionnement du nouveau produit sur le marché français.
Deux versions du produit ont fait l'objet d'un test à domicile en monadique séquentiel (voir chapitre II, section 4.1.1) sur un effectif de 190 foyers (ciblés sur la fréquence de consommation du produit) dans trois villes différentes.
Un questionnaire comprenant de nombreuses questions ouvertes et semi-ouvertes a permis de connaître la réaction spontanée des consommateurs face à ce produit complètement nouveau.
En effet, les résultats de ce test ont permis de connaître :

- les qualités et les défauts du produit
- les façons de consommer le produit
- les façons de préparer le produit
- les raisons de l'achat ou du refus d'achat

Les réponses aux questions ouvertes ont donné des exemples d'items qui seront réutilisés dans un questionnaire fermé.
Ce test a donc servi de matrice à un test de plus grande envergure conduit par le marketing sur l'ensemble de la France.

Pour cette étude, notre méthodologie a donc apporté :

- un coût réduit de test consommateur
- un accès direct au vocabulaire des consommateurs
- une rapidité d'exécution
- une facilité de communication des résultats (auprès des décideurs et du marketing)
- un contrôle total des conditions du test

2.2. Choix d'une formule pour le développement d'un nouveau produit

Le service Développement des Brasseries Kronenbourg désirait connaître la préférence des consommateurs sur deux formules différentes d'un nouveau produit.
Les deux types de produit ont été dégustés à domicile en monadique séquentiel auprès de 199 foyers (ciblés sur la fréquence de consommation du produit) de deux villes différentes.
Le questionnaire comprenait à la fois des questions ouvertes et des questions fermées (voir ANNEXE). Les dernières ont permis de déterminer le pourcentage de préférence pour les deux sortes de produit, ainsi que de réaliser des profils de préférences.
Les questions ouvertes ont quant à elles pu donner les raisons des préférences ou de la non préférence.
Il est important de remarquer que les résultats issus des deux types de questions ont abouti à la même conclusion à savoir la préférence justifiée d'une des deux formules.

Pour cette étude, notre méthodologie a donc apporté :

- des résultats allant dans le même sens que les résultats aux questions fermées
- une confirmation dans la préférence d'un produit
- une description spontanée de cette préférence

2.3. Amélioration d'un produit existant

Ce dernier exemple concerne l'amélioration de la formule aromatique d'un produit de la branche Biscuits du groupe DANONE.
Plus de cent consommateurs ciblés ont dégusté deux produits, représentés par un témoin et un essai à domicile en monadique séquentiel dans un ordre aléatoire.
Deux types de questions ouvertes nous ont aidé à connaître les qualités et les défauts des deux recettes.
Les mêmes données ont été à la fois traitées de façon manuelle, avec le logiciel LEXICO et avec notre méthodologie.
Nous avons d'une part observé davantage de précision avec les traitements automatiques (en plus de la simplicité et rapidité d'exécution). D'autre part, l'absence de codage et la difficulté d'interprétation des résultats en provenance de LEXICO ont conduit à une qualité d'information moins riche qu'avec notre méthodologie.
Néanmoins les résultats sont tous allés dans le même sens avec trois types de traitement.

Pour cette étude, notre méthodologie a donc apporté :

- des résultats confortés par d'autres types de traitement des données textuelles et plus riches d'information
- une description spontanée des points forts et faibles d'une nouvelle formulation de produit
- une interprétation aisée des résultats

3. Principales avancées de la thèse

Pour aborder ce projet de recherche nous avons dû acquérir des connaissances et des compétences multiples en analyse sensorielle, collecte des données consommateurs et traitement des données textuelles. C'est grâce à la rencontre et à la maîtrise de ces différentes disciplines que nous avons pu mettre en place de nouvelles approches.

3.1. Connaissance des méthodes textuelles

Une recherche bibliographique enrichie par diverses informations recueillies lors de rencontres[40] et congrès[41] nous a permis d'acquérir une solide connaissance des méthodes de traitement des données textuelles.
Ceci a non seulement apporté une vision globale des dernières avancées dans le domaine, mais cela a également donné l'occasion de confronter nos travaux avec les ceux d'autres chercheurs. Cela nous a enfin assuré de la nature innovante et prometteuse de notre démarche[42].
La particularité de cette dernière provient de la confrontation d'idées en provenance de deux domaines différents. Aussi, nous avons pu progresser en analyse sensorielle grâce aux connaissances en lexicométrie et en bibliométrie. De la même façon, l'approche des commentaires libres de consommateurs a apporté des éléments nouveaux en traitement des données textuelles.

3.2. Nouvelles approches

Plusieurs nouveaux concepts découlent de ce travail de thèse :

Adaptation des paramètres de calcul

Devant la complexité de la caractérisation sensorielle d'un produit et la particularité du vocabulaire du consommateur, nous nous sommes principalement intéressés aux mots les plus co-cités dans la dégustation d'un même produit. Pour arriver à un résultat concluant, nous nous sommes inspirés des précédents travaux (section 2.1.3. du chapitre II) pour déterminer les unités de segmentation c'est-à-dire l'ensemble des éléments à associer.
Enfin, une comparaison des indices d'association existants dans la littérature nous a permis de choisir le mode de calcul le plus approprié pour nos données.

Réduction du vocabulaire

Afin de manipuler des données homogènes et représentatives statistiquement au niveau des valeurs de fréquences, nous avons mis en place une véritable norme de dépouillement (section 4.2. du chapitre II). Cette dernière est spécifique pour chaque produit alimentaire testé. Elle reste évolutive et réutilisable.
Elle permet de s'affranchir des variations d'interprétations trop souvent rencontrées lors du postcodage manuel.

[40] Avec M. REINERT, A. SALEM, L. LEBART,
[41] TALN, JADT
[42] 4 communications et 2 posters (voir bibliographie)

Pondération des notions

La méthode de codage abordée dans notre démarche est tout à fait nouvelle en traitement de l'information puisqu'elle a été développée spécifiquement pour les commentaires libres de consommateurs.
En effet, le fait de pouvoir associer la nuance quantitative à un descripteur n'avait jamais été réalisé. Cette étape est pourtant indispensable lors de l'interprétation des résultats : si un consommateur annonce qu'il ressent de l'amertume, nous nous demandons tout de suite si elle est perçue de façon forte ou de façon faible.

Représentation graphique

De la même façon, la représentation graphique a été mise en place sur un schéma précis pour répondre à plusieurs besoins particuliers (voir chapitre II, section 3.3 et 4.3). Nous avons choisi d'agir en conséquence car aucune des solutions explorées ne correspondait à notre attente.

Méthode automatique

Ces nouveaux concepts ont été concrétisés par des applications informatiques qui permettent de réaliser des études de manière automatique. Ceci nous permet de gagner du temps et de la fiabilité.

4. Perspectives

Le travail de cette thèse à permis, nous venons de le voir, de répondre à plusieurs questions qui restaient en suspens pour le Groupe DANONE. Répondre aux questions n'est pas une finalité en soi. Il ne faudrait surtout pas oublier de parler des axes de recherche qui, faute de temps n'ont pu être entièrement traités ou encore ceux qui découlent des travaux que nous avons réalisés.

4.1. En recherche

Traitement du langage naturel

Nous nous sommes intéressés au traitement du langage naturel au début du projet (voir section 2.1.2 du chapitre II). La démarche nous a paru tout à fait intéressante et prometteuse. Cependant, elle n'était pas à ce moment là en adéquation avec nos budgets et contraintes de temps.
Le traitement du langage naturel mérite désormais toute notre attention pour envisager la réalisation d'un codage plus rapide, demandant moins d'investissement humain.

Saisie vocale

La recherche de nouveaux modes de recueil des commentaires libres de consommateurs nous a également interpellée en début de thèse. Des logiciels de saisie vocale automatique ont été testés dans l'espoir de rendre la saisie plus simple, plus rapide, plus intelligente. Pour le consommateur, les contraintes d'écriture seraient écartées, il pourrait s'exprimer tout à fait librement. Nous aurions le mode de questionnement de l'entretien semi-directif avec le mode de collecte d'un test sur ordinateur.
Malheureusement, plusieurs raisons nous ont incitées à renoncer à l'utilisation de ce logiciel :

- la période de paramétrage est beaucoup trop longue pour être réalisée par le consommateur (2 heures).
- les résultats sont fortement dépendants de ce paramétrage.
- le logiciel est très sensible aux différences de prononciation, aux accents, aux bruits annexes.
- le taux de réussite est extrêmement faible (48,8 % en moyenne). Les phrases saisies sont totalement incompréhensibles. Un grand nombre de mots sont totalement différents, quelques-uns sont phonétiquement semblables.
- il n'y a pas d'amélioration au cours des différents essais.
- au problème de la mauvaise reconnaissance des mots dictés, il faut également ajouter une mauvaise reconnaissance de la ponctuation, donc de certaines commandes.
- il existe une possibilité de correction du texte erroné. Mais ceci augmenterait énormément la durée totale de saisie !

Cependant, l'évolution de ce type de produit mérite d'être surveillée de près. Aujourd'hui le même type d'outil s'est beaucoup perfectionné.

Lecture hypertextuelle

La représentation graphique des réseaux de connexion nous a incités à pousser notre réflexion plus en avant sur le plan de la lecture des résultats.
D'autre part, l'utilisation de plus en plus intense des ordinateurs comme outil de travail, de l'Internet et des logiciels de CD ROM, nous fait évoluer vers une nouvelle façon de lire les informations.
Par extrapolation et homologie des idées, nous avons pensé que nous pouvions associer des idées de la même manière que nous le faisions avec les mots.
La construction d'un document en hypertexte offre une certaine souplesse au lecteur (voir section 2.3. du chapitre II). C'est une nouvelle approche de la lecture, totalement différente de la lecture classique sur papier. Elle est à la fois plus rapide et plus proche des résultats puisque les liens hypertextes permettent de réaliser un raccourci. Il n'y a plus de sous entendu : c'est une lecture plus profonde et non limitée dans l'espace. C'est sans doute le mode de lecture de demain.
Poussé par l'initiative du Professeur Henri DOU, nous avons voulu mettre en pratique cette réflexion en proposant ce manuscrit de thèse en version CD ROM.

La lecture du document numérique en format *PDF*[43] pourra se réaliser suivant deux axes :

- des parcours colorés en fonction d'une thématique spécifique :
  - veille
  - analyse sensorielle
  - traitement des données textuelles
- navigation dans le sommaire et les références bibliographiques

4.2. En industrie

Questionnaire semi-ouvert

Actuellement, nous pouvons considérer que nous sommes en mesure d'observer le vocabulaire du consommateur.
En effet, nous savons d'une part questionner les consommateurs d'une façon simple en leur permettant de s'exprimer librement. D'autre part, nous sommes capables de traiter ces réponses libres.

L'information résultante est spontanée et indique de façon non orientée, la représentation du produit chez le consommateur.

Pourtant, il existe des points qui méritent réflexion :

- certaines notions sont confuses. EXEMPLE : le désaltérant, le rafraîchissant,...
- le taux de citation est assez faible dans le cas des tests de routine (réalisé sur un petit échantillon de consommateurs).
- le consommateur répond plus facilement aux questions fermées

La fermeture complète des questions n'est pas à envisager comme mode de questionnement, puisque nous souhaitons avant tout garder la spontanéité dans les réponses.
Une solution serait de mettre en place une pré-orientation des questions. Le traitement des questions ouvertes va nous permettre de proposer un éventail assez large de termes aux consommateurs. Cet ensemble de termes représente l'univers du produit dans l'expression du consommateur.
C'est donc dans ce sens que nous envisageons un questionnaire pré-orienté. D'autre part, nous n'excluons pas la possibilité pour le consommateur d'inscrire un nouveau terme non présent dans la liste.
Le consommateur peut donc sélectionner les termes qui lui semblent convenir pour décrire le produit qu'il vient de déguster.
Ce questionnaire doit pouvoir s'adapter à chaque produit alimentaire. Il est donc nécessaire de pouvoir le paramétrer au niveau de la liste de termes, de la pondération et des associations entre les termes.

[43] Portable Document Format d'Adobe™

Ce type de questionnaire a été mis au point en fin de thèse et testé auprès d'un petit groupe de consommateurs. Il a été conçu en page HTML pour l'inscrire dans un projet de construction d'une borne interactive et d'un serveur Internet.

Questionnaire interactif

Cette idée découle de la précédente dans la mesure où le questionnaire interactif doit être construit avec des questions fermées ou semi-ouvertes. Il s'agirait de créer un nouveau mode de récupération des sondages consommateurs avec les technologies issues de l'Internet dans le but de :

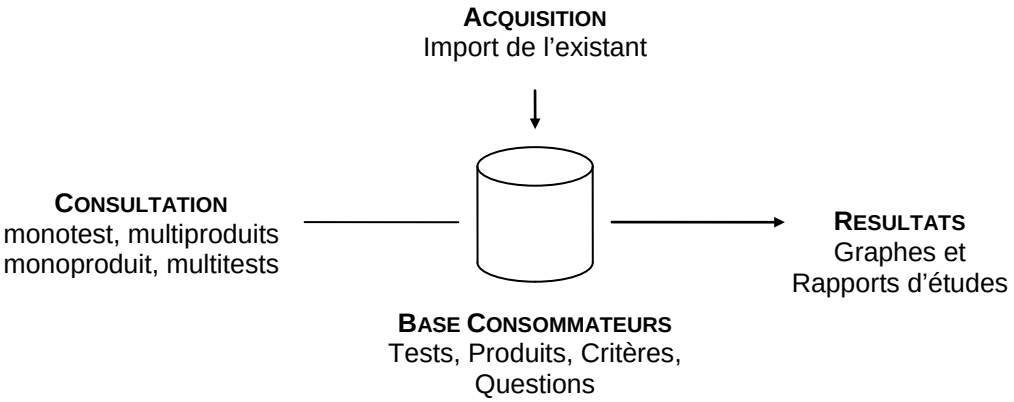
- faciliter les tests consommateurs au niveau de la préparation du test, de la saisie des données, du stockage des données
- rendre les consommateurs actifs dans leur participation à l'étude (encouragement aux bonnes réponses)
- diminuer le coût d'un test consommateur
- promouvoir l'image des produits auprès des consommateurs
- suivre l'évolution des techniques de communication

Le système doit être capable de poser des questions supplémentaires au consommateur en fonction des réponses précédentes. Il doit également être interactif pour permettre à la personne de modifier ses réponses. Enfin, les réponses pourront servir d'argumentation sur certains aspects du produit (historique, fabrication, dégustation, …).

Base de données consommateurs

L'ensemble des tests consommateurs représente une masse d'information importante (collecte des données, résultats statistiques, représentations graphiques, rapports d'études). Il est toujours assez facile de s'y retrouver au moment où nous sommes en train de les manipuler. Mais cela devient moins évident lorsque plusieurs mois voire plusieurs années se sont écoulées. Pourtant, il arrive souvent de faire des retours en arrière sur un test donné ne serait-ce que pour établir des comparaisons dans le temps. Les chances de retrouver facilement la bonne information sans y passer trop de temps sont dépendantes de notre système d'archivage. Malheureusement, ce dernier est variable d'un individu à l'autre et pour un même individu d'une période à une autre.
Un moyen plus fiable pour gérer ce type d'information serait une base de données qui permettra à la fois de :

- recueillir les données issues des tests consommateurs et des résultats issus des traitements de ces tests
- rechercher les critères les plus cités pour un produit donné ou encore les produits les plus cités pour un critère donné
- réaliser des comparaisons entre produits
- déterminer les défauts et les qualités des produits



Cette démarche permettra de capitaliser les données consommateurs et de partager les connaissances.

Capitalisation des connaissances

La gestion des informations consommateurs est tout à fait intéressante pour l'industriel. Mais pour l'analyse sensorielle l'idéal serait d'interfacer plusieurs types de données : les informations en provenance des experts sensoriels, les informations physico-chimiques sur le produit et les informations consommateurs (préférences, commentaires).
C'est dans cet esprit que l'équipe des Sciences du goût du TEPRAL a envisagé de développer une base de connaissances regroupant l'ensemble de ces informations sur la bière.
Cette dernière permettra de consulter le vocabulaire de dégustation des bières sous différents angles :

- celui de l'expert pour mesurer les caractéristiques sensorielles du produit
- celui du produit pour mesurer les caractéristiques physico-chimiques
- celui du consommateur pour mesurer les préférences et observer les caractéristiques de son vocabulaire

Cet outil permettra de mieux appréhender par exemple, quelles notions sensorielles, physico-chimiques et lexicales le terme rafraîchissant peut regrouper.

Formation sur le produit

Actuellement, les brasseurs se rendent compte que l'image de la bière est assez ternie auprès du consommateur. Aussi, une réflexion générale consiste à rechercher un moyen de redorer le blason de ce produit souvent associé aux côtés négatifs comme l'alcoolisme, la violence, les accidents de la route, …

L'étude du vocabulaire de la bière du consommateur met en évidence un corpus de petite taille. Celui-ci est bien moins développé que le vocabulaire moyen du vin.

Une des idées serait d'objectiver l'image culturelle du produit par une mesure de la richesse du vocabulaire employé spontanément en ayant recours à la communication ou à la formation sur le produit.
En effet, nous partons du principe du phénomène de mode : plus les gens parleront du produit, plus son image de marque en sera bonifiée.

4.3. Autres applications

La méthode que nous avons mise en place n'a pas été uniquement appliquée aux commentaires libres de consommateurs.
Nous avons pu en effet, réaliser des études sur d'autres types de données. Il s'agissait bien sûr toujours de données textuelles :

- champs titre de références bibliographiques de brevets
- vocabulaire du fruité des experts sensoriels en bière

Ces deux approches ont révélé des résultats intéressant et assez prometteurs pour envisager d'autres applications dans la même lignée. Parmi elles, nous pensons entre autres :

- aux réclamations clients
- à des entretiens semi-directifs de consommateurs
- à des rapports internes
- à des séquences d'ADN et de protéines

REFERENCES BIBLIOGRAPHIQUES

1. [ASU91]
A.S.U.
Techniques d'analyse et de contrôle dans les industries agro-alimentaires
Lavoisier - Tec & Doc (2° édition), Paris, Volume 2, 1991, p381-449

2. [ASU92]
A.S.U.
La qualité de l'information dans les enquêtes
Dunod, Paris, 1992, 549p

3. [ABEI94]
ABEILLE A., GODARD D.
The complementation of french auxiliaries
UFRL et CNRS, Université de Paris 7, 1994

4. [AFN95a]
AFNOR
Analyse sensorielle- Vocabulaire.
NF ISO 5492 (Indice de classement : V 00 150), 1995, 27-51

5. [AFN95b]
AFNOR
Contrôle de la qualité des produits alimentaires. Analyse sensorielle.
AFNOR, 5° édition, 1995, 400p

6. [ANTO96]
ANTONI M. H.
Text Navigator
ECAM - IBM, SERV/TN/200696, 1996, 12p

7. [AUCO91]
AUCOUTURIER A.-L., BEAUDOUIN V., BLOT I., FAIVRE D., LAHLOU S., MICHEAU J.
Nature et traitement statistique des données textuelles : Réflexions méthodologiques
Cahiers de recherche, CREDOC, n°24, 1991, 53p

8. [BASB87]
BASBERG B.
Patents and the measurement of technological change : a survey of the literature
Research Policy n°16, 1987

9. [BEA93a]
BEAUDOUIN V., BOISBOUVIER N., HEBEL P., LITMAN S., RACAUD T.
L'analyse lexicale : outil d'exploration des représentations; Résultats illustratifs (annexe au cahier de rechercher n°48)
Cahiers de recherche, CREDOC, n°48Bis, 1993, 175p

10. [BEA93b]
BEAUDOUIN V., LAHLOU S.
Réponse à une question ouverte : incidence du mode de questionnement
JADT, Montpellier, 1993, p133-145

11. [BEAU94]
BEAUDOUIN V., HEBEL P.
Avancées en analyse lexicale
Cahiers de recherche, CREDOC, n°61, 1994, 104p

12. [BEAU95]
BEAUDOUIN V.
Analyse textuelle et structures narratives de récits
Cahiers de recherche, CREDOC, n°82, 1995, 42p

13. [BEAU96]
BEAULIEU Y., BERNARD F., FORTIN J.
L'entreprise agroalimentaire : assurer la croissance
Les éditions du monde alimentaire Inc. St-Jean sur Richelieu (Québec), 1996, 346p

14. [BECU93]
BECUE M., PEIRO R.
Les quasi-segments pour une classification automatique de réponses ouvertes
JADT, Montpellier, 1993, p411-423

15. [BEN73a]
BENZECRI J. P.
L’analyse des données ; Taxinomie
Dunod, tome 1, 1973

16. [BEN73b]
BENZECRI J. P.
L’analyse des données ; L’analyse des correspondances
Dunod, tome 2, 1973

17. [BERN88]
BERNET C., DUBROCARD M., LABBE D., BRAINERD B., HOLMES D. I., SERANT D., BRUNET E., HUBERT P. THOIRON P.
Etudes sur la richesse et la structure lexicales
Edition Slatkine Champion, Paris, 1988,

18. [BERT77]
BERTIN J.
Le graphique et le traitement graphique de l'information
Flammarion, Paris, 1977, 277p

19. [BOUC93]
BOUCHE R., GERMAIN N.
Bibliométrie, infométrie et analyse automatique de documents écrits
Les systèmes d'information élaborés, Ile Rousse, 1993, p352-365

20. [BOUT96]
BOUTIN E., QUONIAM L., ROSTAING H., DUMAS P.
Traitement de l'information : analyse des données classiques versus analyse réseau. Un cas d'application : la bibliométrie
Inforcom. Université Stendhal de Grenoble: Université Lille III, 1996, p571-587

21. [BROU93]
BROUSTAIL J., FRERY F.
Le management stratégique de l'innovation
Dalloz, Paris, 1993

22. [BRUN93]
BRUNET E.
Une hypertexte statistique : Hyperbase
JADT, Montpellier, 1993, p1-16

23. [BURE89]
BUREAU G., MULTON J. L.
L'emballage des denrées alimentaires de grande consommation
Lavoisier Tec & Doc, collection Sciences & Techniques Agroalimentaires, 1989, 729p

24. [CAIL76]
CAILLEZ F., PAGES J.P.
Introduction à l'analyse des données
Société de Mathématiques Appliquées et de Sciences Humaines, BURO, 1976, 616p

25. [CALL93]
CALLON M., COURTIAL J.-P., PENAN H.
La scientométrie
Edition Presses universitaires de France, Paris, 1993, 126p

26. [CASE97]
CASES L.
La consommation des ménages en 1996
INSEE Première, n°520, 1997, 4p

27. [CEMA98]
CEMAGREF
Institut de recherche pour l'ingénierie de l'agriculture et de l'environnement
http://www.cemagref.fr/, mars 1998

28. [CHAU92]
CHAUMIER J., DEJEAN M.
L'indexation assistée par ordinateur ; Principes et méthodes
Documentaliste, Sciences de l'information, volume 29, n°1, 1992, p3-6

29. [CHEM96]
CHEMDATA
TEXTO pour WINDOWS ; Manuel de référence
Version 6.0., 1996

30. [CIBO82]
CIBOIS P.
Tri-deux : une méthode post-factorielle de dépouillement d'enquête
L'année sociologique,n°32, 1982, p62-80

31. [CORE94]
CORET A., MENON B., SCHIBLER D., TERRASSE C.
Un système d'indexation structurée à l'INIST ; Bilan d'une étude préalable
Documentaliste, Sciences de l'information, volume 31, n°1, 1994, p148-158

32. [CNEV98]
CNEVA
Centre National d'Etudes Vétérinaires et Alimentaires
http://194.51.251.1/srpc/cnevahtm/sommaire.htm, mars 1998

33. CNRS98]
CNRS
Institut Européen des Sciences du Goût et des Comportements Alimentaires
http://www.infobiogen.fr/SDV/cesglabo.html, mars 1998

34. [COHE96]
COHEN-SOLAL M
1995 : une année plutôt bonne pour les industries agro-alimentaires
INSEE Première, n°468, 1996, 4p

35. [COHE97]
COHEN-SOLAL M.
Les industries agro-alimentaires en 1996 : Croissance confirmée malgré la crise de la viande bovine
INSEE Première, n°528, 1997, 4p

36. [COMM94]
COMMISSARIAT GENERAL DU PLAN
Rapport du Groupe " Intelligence économique et stratégie des entreprises "
La Documentation française, Paris, 1994, 213p

37. [CONF97]
CONFLAND D.
Economie de l'information spécialisée
ADBS, Paris, 1997, 347p

38. [COUR76]
COURRIER Y.
Analyse et langage documentaire
Documentaliste, vol13, n°5-6, 1976, p 178-189

39. [COUR94]
COURTIAL J.-P., POCHON J., VILAIN C.
L'étude d'un concept nouveau à partir de réseaux de mots-clés; Application à la didactique des sciences
Documentaliste - Sciences de l'information, vol. 31, n°4-5, 1994, p199-204

40. [CRIT93]
CRITON Y., DENEFLE S., JUIN R., QUESADA R., ROUX N.
Trois logiciels, trois interprétations ? Analyse comparative d'un même corpus
JADT, Montpellier, 1993, p103-111

41. [DANO97]
DANONE
Rapport annuel 1996
Groupe DANONE, Paris, 1997, 64p

42. [DANO98]
DANONE
Meet the DANONE Group
http://www.danonegroup.com/Meet_The_Danone_Group/, mars 1998

43. [DESV92]
DESVALS H., DOU H.
La veille technologique
DUNOD, Paris, 1992, 436p

44. [DEVI92]
DEVILLE J. C.
Elements pour une théorie des enquêtes pas quotas ; In La qualité de l'information dans les enquêtes, ASU
DUNOD, 1992, p345-364

45. [DOUH92]
DOU H
Le système d'information lié à la veille technologique
Conférence de l'Association aéronautique et astronautique de France, Management et information : de la synthèse à la décision, Strasbourg, 1992

46. [DOUH95]
DOU H.
La veille technologique et le développement industriel : de la grande entreprise aux PME/PMI
Dunod, Paris, 1995

47. [DUGA80]
DUGAST D.
La statistique lexicale
Editions Slatkine, Genève, 1980, 105p

48. [DUMA93]
DUMAS S., QUONIAM L.
Exploitation de l'enquête des besoins recensés par le CETIM. Utilisation des coefficients de similitude et de dissimilitude
Les systèmes d'information élaborés, Ile Rousse, 1993, p402-417

49. [DUMA94]
DUMAS S.
Développement d'un système de veille stratégique dans un centre technique
Thèse de doctorat, Université d'Aix-Marseille III, 1994, 209p

50. [EINA97]
EINARSSON E., JONES W.
La réforme réglementaire dans l'agro-alimentaire
L'observateur de l'OCDE, n°206, 1997, p23-27

51. [EJER95]
EJERHED E.
Linguistic and computational principles for tagsets
TALN, Marseille, 1995, p34-42

52. [EURO91]
EUROSTAF
Les leaders européens de l'agro-alimentaire face à leurs concurrents américains : le poids de la marque
Paris, 1991

53. [FAOG98]
FAO/GIEWS
Food Outlook ; Global Information and Early Warning System
http://www.fao.org/WAICENT/faoinfo/economic/giews/english/fo/fo9802/httoc.htm, n° 1, Rome, February 1998

54. [FUCH93]
FUCHS C.
Linguistique et traitements automatiques des langues
Hachette, Paris, 1993, p83-104

55. [GAZD95]
GAZDAR G., KLEIN E., PULLUM G., SAG I.
Generalized phrase structure grammar
Cambridge, Harvard University Press, 1995

56. [GDRP95]
GDR-PRC
Le Traitement Automatique du Langage Naturel 1995
Actes de colloque, Marseille, 1995, 263p

57. [GOUT93]
GOUTTAS C., WARNESSON I.
Des outils d'indexation couplés à l'analyse relationnelle pour l'exploitation des données textuelles
JADT, Montpellier, 1993, p271-280

58. [GRAN93]
GRANGE D., LEBART L.
Traitements statistiques des enquêtes
Dunod, Paris, 1993, 255p

59. [GREM87]
GREMY J. P.
Les expériences françaises sur la formulation des questions d'enquête
Revue française de sociologie, XXVIII, 1987, p567-599

60. [GREM92]
GREMY J. P.
La formulation des questions d'enquête : son effet sur les réponses ; In La qualité de l'information dans les enquêtes, ASU
Dunod, 1992, p97-114

61. [GUEL94]
GUELLEC D., KABLA I.
Le brevet : un instrument d'appropriation des innovations technologiques
Economie et Statistique, n°276-3, 1994

62. [HENA73]
HENAULT G. M.
Le comportement du consommateur ; Une approche multidisciplinaire
Les presses de l'Université de Québec, Montréal, 1973

63. [HENA79]
HENAULT G. M.
Le consommateur
Les presses de l'Université de Québec, Montréal, 1979,

64. [HOLL96]
HOLLINGSWORTH P.
Sensory testing and the language of the consumer
Food Technology, 1996, p65-69

65. [IFRE98]
IFREMER
Institut français pour la recherche de l'exploitation de la mer
http://www.ifremer.fr/, mars 1998

66. [INFO94]
INFORMATION & COMMUNICATION
Infotrans classic pour MS DOS ; manuel d'utilisation
Version 4.0, 1994

67. [INRA98]
INRA
Institut national de la recherche agronomique
http://www.inra.fr/, mars 1998

68. [INSE97]
INSEE
Images économiques des entreprises en 1996
INSEE Résultats, n°589, tome 2, 1997

69. [ISHI87]
ISHII R., O'MAHONY M.
Taste sorting and naming : can taste concepts be misrepresented by traditional psychophysical labelling systems ?
Chemicals senses, 12(1), 1987, p37-51

70. [ISSA92]
ISSANCHOU S., HOSSENLOP J.
Les mesures hédoniques, méthodes, portées et limites" in Plaisir et préférences alimentaires
Polytechnica, 1992, p49-75

71. [JACQ93]
JACQUEMIN C.
A coincidence detection network for spatio-temporal coding : application to nominal composition
Thirteenth International Joint Conference on Artificial Intelligence, Chambéry, Volume 2, 1993, p1346-1351

72. [JACQ94]
JACQUEMIN C.
Optimizing the computational lexicalization of large grammars
ACL, 1994

73. [JAKO94]
JAKOBIAK F.
Le brevet source d'information
Dunod, Paris, 1994, 191p

74. [JUAN86]
JUAN S.
L'ouvert et le fermé dans la pratique du questionnaire ; Analyse comparative et spécificités de l'enquête par correspondance
Revue française de sociologie, XXVII, 1986, p301-316

75. [KABL94]
KABLA I.
Un indicateur de l'innovation : le brevet
Economie et Statistique, n°276-4, 1994

76. [KAPF89]
KAPFERER J. N., THOENIG J.C.
La marque, moteur de la compétitivité des entreprises et de la croissance de l'économie
Mc Graw-Hill, Paris, 1989, 384p

77. [KERI93]
KERIHUEL A.
L'industrie agro-alimentaire et l'innovation
Agreste analyses et études, n°13, 1993, p27-34

78. [LABB]
LABBE D., HUBERT P.
La richesse du vocabulaire
Colloque de l'ALLC-ACH, Paris

79. [LABB90]
LABBE D.
Normes de saisie et de dépouillement des textes politiques
Cahier du C.E.R.A.T., n°7, 1990, 135p

80. [LABB92]
LABBE D.
Normalisation des textes et statistique lexicale sur MacIntosh
CERAT, version préliminaire, 1992, 26p

81. [LABB98]
LABBE D., HUBERT P.
La structure du vocabulaire du Général De Gaulle
JADT, Rome, volume 2, 1995, p165-174

82. [LAFO80]
LAFON P.
Sur la variabilité de la fréquence des formes dans un corpus
Mots n°1, Presses de la fondation nationale des sciences politiques, 1980, p127-165

83. [LAFO81]
LAFON P.
Analyse lexicométrique et recherche des cooccurrences
Mots n°3, Presses de la fondation nationale des sciences politiques, 1981, p95-148

84. [LAFO83]
LAFON P., SALEM A.
L'inventaire des segments répétés d'un texte
Mots n°6, Presses de la fondation nationale des sciences politiques, 1983, p161-177

85. [LAFO85]
LAFON P., LEFEVRE J., SALEM A., TOURNIER M.
Le Machinal. Principes d'enregistrement informatique des textes
Paris, Klincksieck, 1985,

86. [LAFO92]
LAFOUGE T., QUONIAM L.
Les distributions bibliométriques
Revue française de bibliométrie, volume 9, 1992, p128-138

87. [LAHL92]
LAHLOU S.
Si/alors : "Bien manger" ?. Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus consititué des associations libres de 2000 individus
Cahiers de recherche, CREDOC, n°34, Avril 1992, 161p

88. [LAHL93]
LAHLOU S., BEAUDOUIN V.
L'analyse lexicale : outil d'exploration des représentations
Cahiers de recherche, CREDOC, n°48, 1993, 145p

89. [LAHL95]
LAHLOU S.
Penser manger; Les représentations sociales de la l'alimentation
Thèse de doctorat nouveau régime, Ecole des Hautes Etudes en sciences sociales, 3 Tomes, 1995, 448p

90. [LAUR97]
LAURI P., ZIEGELBAUM H.
Réflexion sur l'évolution de la profession de Veilleur
Les systèmes d'information élaborés, Ile Rousse, 1997

91. [LEBR96]
LE BRIS F.
Mondialisation industrielle : le rôle des filiales de commerce de gros des groupes étrangers
INSEE Première, n°485, 1996, 4p

92. [LECR90]
LE CROSNIER H.
Système d'accès à des ressources documentaires - vers des antéserveurs intelligents
Thèse de doctorat, Université Aix-Marseille III, 1990, 355p

93. [LEBA88]
LEBART L; SALEM A.
Analyse statistique des données textuelles
Dunod, Paris, 1988, 209p

94. [LEB93a]
LEBART L;, MORINEAU A., BECUE M., HAEUSLER L.
Introduction à SPAD.T intégré, Version 1.5 PC
CISIA, Saint-Mandé, 1993, 130p

95. [LEB93b]
LEBART L.
Traitement des questions ouvertes ; In Traitements statistiques des enquêtes
Dunod, Paris, 1993, p227-246

96. [LEBA94]
LEBART L;, SALEM A.
Statistique textuelle
Dunod, Paris, 1994, 342p

97. [LEGE84]
LEGENDRE L., LEGENDRE P.
Ecologie numérique
Masson, Presses de l'Université du Québec, tome 1 et 2, 1984

98. [LELO98]
LELOUP C.
Moteurs d'indexation et de recherche ; Environnement client-serveur, Internet et Intranet
Eyrolles, Paris, 1998, 277p

99. [LION91]
LION S.
Construction d'un corpus et perte d'information en analyse lexicale
Cahiers de recherche, CREDOC, n°13, 1991, 61p

100. [MALA86]
MALASSIS L.
Economie agro-alimentaire
Cujas, Tome 1 et 3, 1986

101. [MARC81]
MARCOTORCHINO F., MICHAUD P
Agrégation de similarités en classification automatique
Etude IBM, n°F-012, Centre scientifique de Paris, Janvier 1981

102. [MART89]
MARTINET B., RIBAULT J. M.
La veille technologique concurrentielle et commerciale
Editions d'organisation, Paris, 1989, 300p

103. [MART93]
MARTIN N.
Exploration d'un espace de perceptions et d'un espace de préférences; Recherche d'optima en formulation sensorielle
Thèse de doctorat, Ecole Nationale Supérieure des Industries Agricoles et Alimentaires, 1993, 223p

104. [MART94]
MARTIN N., ROGEAUX M.
Etude par analyse textuelle de commentaires de consommateurs après dégustation de boisson
Sciences des aliments, n°14, 1994, p265-280

105. [MARZ96]
MARZLOFF B., BELLANGER F.
Les nouveaux territoires du marketing ; enquête sur les réponses des médias au géomarketing et au marketing relationnel
Editions Liaisons, Paris, 1996, 287p

106. [MEIL79]
MEILGAARD M. C., CIVILLE G.V., CARR B.T.
Beer flavour terminology
J. Amer. Soc. Brew. Chem., n°37, 1979, p47-59

107. [MICH88]
MICHELET B.
L'analyse des associations
Thèse de doctorat, Université de Paris VII, 1988, 407p

108. [MILL94]
MILLSTONE E.
Regulation, innovation and public welfare : The example of the food industry
Technology Analysis & Strategic Management, Volume 6, n° 3, 1994, p329-340

109. [MINI97]
MINISTERE DE L'EDUCATION NATIONALE, DE LA RECHERCHE ET DE LA TECHNOLOGIE
La recherche : une ambition pour la France
http://www.mesr.fr/gouv/jaune/index.htm, 1997

110. [MINI98]
MINISTERE DE L'AGRICULTURE ET DE LA PECHE
La politique française de l'alimentation
http://www.agriculture.gouv.fr/alimentation/dgal.stm, mars 1998

111. [MONC95]
MONCEAU C.
30 ans d'échanges agro-alimentaire français : 1961-1990
INSEE Résultats, n°417, 1995

112. [MONC97]
MONCEAU C., DE PERETTI G.
Le commerce extérieur agro-alimentaire de la France et de l'Union Européenne
INSEE Première, n°540, 1997, 4p

113. [MOSC90]
MOSCAROLA J.
Enquêtes et analyse de données
VUIBERT GESTION, Paris, 1990, 307p

114. [MUL92a]
MULLER C.
Principes et méthodes de statistiques lexicales
Editions Champion, Paris, réimpression de l'édition Hachette de 1977, 1992, 205p

115. [MUL92b]
MULLER C.
Initiation aux méthodes de la statistique linguistique
Editions Champion, Paris, réimpression de l'éd. Hachette de 1973, 1992, 185p

116. [MULT91]
MULTON J. L.
Techniques d'analyse et de contrôle dans les industries agroalimentaires ; Le contrôle de qualité : principes généraux et aspects législatifs
Lavoisier Tec & Doc, collection Sciences & Techniques Agroalimentaires, volume 1, 2° édition, 1991, 365p

117. [NEFU89]
NEFUSSI J.
Les industries agro-alimentaires
Que sais-je ? , PUF, Paris, 1989

118. [NIEL95]
NIEL X
Les industries agro-alimentaires en 1994
INSEE Résultats, n°413, 1995

119. [PIBA98]
PIBAROT A., PICARD J., LABBE D.
Les syntagmes répétés dans l’analyse des commentaires libres
JADT, Nice, 1998, p507-515

120. [PETR88]
PETROF J. V.
Comportement du consommateur et marketing
Presse de l’Université de Laval, 4° édition, 1988

121. [PFIR97]
PFIRSCH J.V.
La saveur des sociétés. Sociologie des goût alimentaires en France et en Allemagne
Presses universitaires de Rennes, Collection "le sens social", Rennes, 1997, 206p

122. [PLAN95]
PLANTY B.
Pour un pilotage raisonné des marques
Humanisme & Entreprise, n°31, 1995, p73-78

123. [POLL94]
POLLARD C., SAG I. A.
Head-driven phrase stracture grammar
CSLI, Standford, The University of Chicago Press, 1994

124. [QUON88]
QUONIAM L.
Bibliométrie informatisée et information stratégique
Thèse de doctorat, Université d'Aix-Marseille III, 1988

125. [QUON92]
QUONIAM L.
Bibliométrie sur des références bibliographiques : méthodologie ; In DESVALS H, DOU H, 1992. La veille technologique
DUNOD, Paris, 1992, p244-262

126. [REIN83]
REINERT M.
Une méthode de classification descendante hiérarchique : application lexicale par contexte (C.D.H. lexicale)
Les cahiers de l'analyse des données, Vol. VIII, 1983, n°2, 1983, p187-198

127. [REIN86]
REINERT M.
Un logiciel d'analyse lexicale : ALCESTE
Les cahiers de l'analyse des données, Vol. XI, 1986, n°4, 1986, p471-481

128. [REIN93]
REINERT M.
Les " mondes lexicaux " et leur " logique " à travers l'analyse statistique d'un corpus de récits de cauchemars
Langage et société n°66, 1993

129. [ROEH]
ROEHRICH G., VALETTE-FLORENCE P.
Apport des chaînages cognitifs à la segmentation des marchés
p479-498

130. [ROGE96]
ROGEAUX M., ZIEGELBAUM H.
Comment DANONE prend-il en compte les commentaires sensoriels des consommateurs ?
AGORAL, Lavoisier TEC&DOC, 1996, p139-147

131. [ROST93]
ROSTAING H.
Veille technologique et bibliométrie : concepts, outils, applications
Thèse de doctorat, Université d'Aix-Marseille III, 1993

132. [ROST96]
ROSTAING H.
La bibliométrie et ses techniques
Sciences de la société, Collection " outils et méthodes ", 1996, 131p

133. [ROST98]
ROSTAING H., ZIEGELBAUM H., BOUTIN E., ROGEAUX M., QUONIAM L.
Analyse de commentaires libres par la technique des réseaux de segments
JADT, Nice, 1998, p697-704

134. [SALE93]
SALEM A.
Méthodes de la statistique textuelle
Thèse d'état de l'Université Sorbonne Nouvelle (Paris 3), 3 volumes, 1993

135. [SAUV91]
SAUVAGEOT F.
Techniques d'analyse et de contrôle dans les industries agroalimentaires ; Principes et techniques d'analyse
Lavoisier Tec & Doc, collection Sciences & Techniques Agroalimentaires, volume 2, 2° édition, 1991, p382-448

136. [SCHL92]
SCHLICH P., MCEWAN J. A.
La cartographie des préférences ; un outil statistique pour l'industrie agroalimentaire
Sciences des Aliments n°12, 1992, p339-355

137. [SCHL93]
SCHLICH P.
Uses of change-over designs and repeated mesurements in sensory and consumer studies
Food Quality and Preference, 4, 1993, p223-235

138. [SECO97]
SECODIP
Le marketing book
SECODIP, Chambourcy, 1997, 448p

139. [SFBA95]
SFBA
Les systèmes d'information élaborés
Actes de colloque, Ile Rousse, 1995, 601p

140. [SSHA90]
SSHA
Evaluation sensorielle ; Manuel méthodologique
Apria, TEC&DOC, 1990, 328p

141. [SYLV92]
SYLVANDER B., LASSAULT B.
L'enjeu économique de la qualité sur les marchés des produits agro-alimentaires
INRA-ESRT - 92-01 , 1992, 45p

142. [TEIL91]
TEIL G.
CANDIDE, un outil de sociologie assistée par ordinateur pour l'analyse quali-quantitative de gros corpus de textes
Thèse de doctorat de l'Ecole des Mines de Paris, 1991, 355p

143. [TEI92a]
TEIL G.
Décrire les goûts des fromages : des consommateurs aux experts
Economie et sociologie rurales, GRIGNON, INRA,. n°7, vol. 1 - Le vocabulaire de la dégustation, 1992, 109p

144. [TEI92b]
TEIL G.
Des occurrences à la sémantique : le réseau de mots associés
Colloque Intelligence Artificiel et Textes, Jussieu, 1992

145. [TEI94a]
TEIL G.
Le vocabulaire des dégustateurs "amateurs" de fromages
Economie et sociologie rurales, GRIGNON, INRA,. n°17, 4 vol., 1994

146. [TEI94b]
TEIL G.
Les commentaires de dégustation des consommateurs : une mine encore à explorer
1994, 37p

147. [URSO97]
URSO P.
Mesurer les sensations pour concevoir des produits attrayants
Technologies Internationales n°31, 1997, p36-39

148. [VALM95]
VALMIER J.
Evolution de la distribution
Humanisme & Entreprise, n°40, 1995, p93-106

149. [VANV94]
VAN VRACEM P., JANSSENS-UMFLAT M.
Comportement du consommateur ; Facteurs d'influence externe
De Boeck, Bruxelles, 1994, 365p

150. [VERM95]
VERMERSCH M.
L'industrie tâtonne dans l'analyse sensorielle
L'usine Nouvelle n°2523, 1995

151. [VINC91]
VINCK D.
Gestion de la recherche ; Nouveaux problèmes, nouveaux outils
De Boeck, Bruxelles, 1991, 567p

152. [YVON90]
YVON F.
L'analyse lexicale appliquée à des données d'enquêtes : état des lieux
Cahiers de recherche, CREDOC, n°5, 1990, 66p

153. [ZIEG96]
ZIEGELBAUM H., ROGEAUX M.
Trois exemples de traitement et d'utilisation des commentaires libres de consommateurs
AGORAL 96, Lavoisier TEC&DOC, 1996, p139-147

154. [ZIEG97]
ZIEGELBAUM H., ROGEAUX M., ROSTAING H.
Une méthode de traitement automatique des questions ouvertes
Les systèmes d'information élaborés, Ile Rousse, 1997

155. [ZIEG98]
ZIEGELBAUM H., ROSTAING H., ROGEAUX M.
Utilisation des questions ouvertes dans les tests consommateurs en analyse sensorielle
JADT, Nice, 1998, p649-657

156. [ZIPF49]
ZIPF G. K.
Human behaviour and the principle of least effort
Editions Addison Wesley, 1949, 257p

ANNEXES

ANNEXE 1 : PLAN D'EXPERIENCE

N°	Paire	Jour 1	Jour 2	Jour 3	Jour 4	Jour 5	Jour 6	Jour 7	Jour 8	Jour 9	Jour 10	
20	A4-B2	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
20	A4-B2	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
20	A4-B2	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
21	A4-B3	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
21	A4-B3	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
21	A4-B3	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
22	A4-B4	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
22	A4-B4	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
22	A4-B4	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
23	A4-B5	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
23	A4-B5	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
23	A4-B5	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
24	A4-B6	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
24	A4-B6	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
24	A4-B6	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
25	A5-B1	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
25	A5-B1	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
25	A5-B1	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
26	A5-B2	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
26	A5-B2	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
26	A5-B2	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
27	A5-B3	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
27	A5-B3	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
27	A5-B3	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
28	A5-B4	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
28	A5-B4	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
28	A5-B4	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
29	A5-B5	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
29	A5-B5	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
29	A5-B5	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
30	A5-B6	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
30	A5-B6	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
30	A5-B6	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
31	A6-B1	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
31	A6-B1	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
31	A6-B1	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
32	A6-B2	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
32	A6-B2	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
32	A6-B2	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
33	A6-B3	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
33	A6-B3	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
33	A6-B3	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
34	A6-B4	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
34	A6-B4	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
34	A6-B4	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
35	A6-B5	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
35	A6-B5	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
35	A6-B5	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
36	A6-B6	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	H<35	10
36	A6-B6	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	H>35	10
36	A6-B6	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	F<35	F>35	10
		108	108	108	108	108	108	108	108	108	108	

ANNEXE 2 : EXEMPLE DE QUESTIONNAIRE CONSOMMATEUR SPECIFIQUE QUESTIONS OUVERTES

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

TEST TEXTUEL CONSOMMATEUR

Bonjour,

vous êtes invités à déguster **deux bières différentes**.

Le présent questionnaire comporte deux parties :

- une partie pour recueillir vos commentaires sur la première bière en pages 2, suivie de quelques renseignements personnels (votre identité, vos habitudes de consommation) en page 3, 4 et 5.
- une partie pour recueillir vos commentaires sur la deuxième bière en page 6.

Afin de bien distinguer les deux produits, vous devez boire un peu d'eau et manger un bout de cracotte.

Les bières que vous allez déguster sont commercialisées sous des marques différentes. Elles peuvent être **proches** ou **différentes** en goût les unes des autres.

Pour ce test, il s'agit de décrire **spontanément** les sensations que vous éprouvez en les dégustant, les remarques que vous pouvez faire sur leurs qualités et leurs défauts et enfin imaginer dans quelles circonstances vous les dégusteriez.

Il est important que vos commentaires restent indépendants pour chacune des deux bières. Ne faites pas de comparaison.

Merci et bonne dégustation....

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

Avez-vous soif ? OUI ❑ NON ❑

Après avoir bu la première bière, donnez une note de 1 à 10 traduisant votre jugement:

❑	❑	❑	❑	❑	❑	❑	❑	❑	❑
1 très mauvais	2	3	4	5	6	7	8	9	10 très bon

Vos commentaires

Quelles sont les sensations que vous avez ressenties en buvant cette bière :
.....
.....
.....
.....
.....
.....

Citez les principales qualités de cette bière :
.....
.....
.....
.....
.....
.....

Citez les principaux défauts de cette bière :
.....
.....
.....
.....
.....
.....

Dans quelles circonstances consommeriez-vous cette bière ?
.....
.....
.....
.....
.....
.....

Buvez un peu d'eau et prenez un morceau de Cracotte...

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

Voici quelques questions pour mieux vous connaître :

1. SEXE : HOMME : ❑ FEMME : ❑

2. AGE : 18 - 24 ans : ❑ 25-34 ans ❑ 35-45 ans ❑ > 45 ans : ❑

3. Consommez - vous de la bière :

❑	❑	❑	❑	❑	❑
1 x / jour	2 à 3 x / semaine	1 x / semaine	2 à 3 x / mois	1 x / mois	plus rarement

4. En général, vous buvez de la bière pour :

vous rafraîchir : ❑ déguster : ❑ les 2 : ❑

5. Quel type de bières consommez-vous et à quelle fréquence : (cochez une case par ligne)

		1x par jour	1 à 2 x par semaine	1 à 2 x par mois	jamais
a) LEGERES OU SANS ALCOOL		❑	❑	❑	❑
b) BLONDES DE LUXE	KRONENBOURG	❑	❑	❑	❑
	33 EXPORT	❑	❑	❑	❑
	KANTERBRAU	❑	❑	❑	❑
c) BLONDES SPECIALES :	1664	❑	❑	❑	❑
	GOLD	❑	❑	❑	❑
	HEINEKEN	❑	❑	❑	❑
d) BELGES (CHIMAY, DUVEL, GUEUSE,...)		❑	❑	❑	❑
e) BRUNES OU AMBREES (1664 BRUNE, GUINNESS, G. KILLIANS,...)		❑	❑	❑	❑
f) AUTRE : (précisez)..... :		❑	❑	❑	❑

6. Vous arrive-t-il de déguster la bière à la pression ?

souvent ❑ de temps en temps ❑ jamais ❑

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

7. Avez-vous mangé quelque chose dans l'heure qui a précèdé ce test ?

OUI ❑ NON ❑ si OUI, qu'avez-vous mangé ? : .....

8. Avez-vous bu quelque chose dans l'heure qui a précèdé ce test ?

OUI ❑ NON ❑ si OUI, qu'avez-vous bu ? : .....

9. Vers quel âge avez-vous commencé à consommer des boissons alcoolisées ?

- ❑ moins de 14 ans
- ❑ de 14 à 16 ans
- ❑ de 16 à 18 ans
- ❑ de 18 à 20 ans
- ❑ plus de 20 ans

10. Quelle boisson alcoolisée avez-vous commencé à boire ?
(choisissez une seule réponse)

- ❑ du vin
- ❑ de la bière
- ❑ de l'alcool fort (type whisky, vodka...)
- ❑ autres. Précisez : .....

11. Dans quelle circonstance la consommiez-vous ?

- ❑ à table, en famille
- ❑ à table, entre amis
- ❑ au café
- ❑ en soirée, pendant une fête

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

12. Vers quel âge avez-vous commencé à consommer de la bière ?

- ❑ moins de 14 ans
- ❑ de 14 à 16 ans
- ❑ de 16 à 18 ans
- ❑ de 18 à 20 ans
- ❑ plus de 20 ans

13. De quelle marque s'agissait-il ?

.....

14. Continuez-vous à en consommer ?

OUI ❑ NON ❑

☞ Le test continu, vous allez déguster une seconde bière.

Ne rien Inscrire ici : date : .....n° paire : ..... n° quest. : .....

Avez-vous soif ? OUI ❑ NON ❑

Après avoir bu la deuxième bière, donnez une note de 1 à 10 traduisant votre jugement:

❑	❑	❑	❑	❑	❑	❑	❑	❑	❑
1 très mauvais	2	3	4	5	6	7	8	9	10 très bon

Vos commentaires

Quelles sont les sensations que vous avez ressenties en buvant cette bière :
.....
.....
.....
.....
.....
.....

Citez les principales qualités de cette bière :
.....
.....
.....
.....
.....
.....

Citez les principaux défauts de cette bière :
.....
.....
.....
.....
.....
.....

Dans quelles circonstances consommeriez-vous cette bière ?
.....
.....
.....
.....
.....
.....

Le test est maintenant terminé, nous vous remercions de votre participation.

ANNEXE 3 : EXEMPLE DE QUESTIONNAIRE CONSOMMATEUR CLASSIQUE

QUESTIONNAIRE USAGES & ATTITUDES

1. Pourriez-vous d'abord me dire si vous aimez ou non les aliments suivants :

	Oui	Non	
Café sans sucre			**(7)**
Café avec du sucre			**(8)**
Chocolat noir			**(9)**
Pamplemousse			**(10)**
Endive braisée			**(11)**
	1	2	

1b. Parmi les boissons suivantes, quelles sont celles qu'il vous arrive de boire ?

	Oui	Non	
Bière			**(12)**
Bière sans alcool			**(13)**
Bière brune			**(14)**
Boisson aux fruits (type Oasis)			**(15)**
Cocktails à base de gin, téquila ou vodka			**(16)**
Cola (Pepsi, Coca)			**(17)**
Eau minérale			**(18)**
Panache			**(19)**
Pastis			**(20)**
Tonics (type Schweppes, Gini)			**(21)**
Vin			**(22)**
Whisky			**(23)**
Cidre			**(24)**
Thé			**(25)**
Boissons à base de thé (comme Liptonic)			**(26)**
Jus de fruits			**(27)**
	1	2	

2. Je vais maintenant vous citer quelques remarques, données par d'autres consommateurs de bière. Pour chacune d'entre elles, pouvez-vous me dire à quel point vous êtes d'accord ?

	Pas du tout d'accord	Plutôt pas d'accord	Ni d'accord ni pas d'accord	Plutôt d'accord	Tout à fait d'accord	
La bière est une boisson que je propose volontiers à mes invités						**(27)**
La bière est une boisson que je bois pour me désaltérer quand j'ai soif						**(28)**
La bière est une boisson que je ne bois que dans certaines occasions						**(29)**
La bière est une boisson dont j'aime explorer la variété des goûts						**(30)**
La bière est une boisson que je bois par plaisir						**(31)**
La bière est une boisson que je bois seulement pour accompagner les autres						**(32)**
La bière est une boisson bon marché						**(33)**
La bière est une boisson qui a un goût trop marqué						**(34)**
Je prends toujours la même marque de bière						**(35)**
Pour moi le prix d'une bière est plus important que sa marque						
Je ne bois de la bière qu'en été						**(37)**
J'aime boire la bière sans verre, directement au goulot						**(38)**
La bière est une boisson à la mode						**(39)**
	1	2	3	4	5	

3. Parmi les occasions que je vais vous citer, dites-moi toutes celles au cours desquelles vous buvez de la bière :

Chez moi pendant les repas	☐	**1**	**(40)**
Chez moi, dans l'après-midi	☐	**2**	
Chez moi, dans la soirée	☐	**3**	
Chez moi quand il y a des invités	☐	**4**	
Chez des amis	☐	**5**	
Au café, au bar dans la journée	☐	**6**	
Le soir, dans un bar à bière	☐	**7**	
Quand je suis seul dans un café ou un bar	☐	**8**	
Dans un café ou un bar avec des amis	☐	**9**	
Uniquement au café, jamais chez moi	☐	**1**	**(41)**
Au restaurant	☐	**2**	
Dans une discothèque, ou une fête ou l'on danse	☐	**3**	
A l'apéritif, juste avant de manger	☐	**4**	
Après un effort ou une activité sportive	☐	**5**	
Pendant un repas léger pris sur le pouce	☐	**6**	
Dans une cafétéria comme Flunch ou un fast food comme MacDonald's	☐	**7**	

4. Quelles sont pour vous, et d'après cette liste (montrer la liste) les 3 qualités essentielles d'une bonne bière ?

		(42)			**(43)**
Plutôt très pétillante	☐	**1**	Plutôt forte en alcool	☐	**1**
Plutôt peu pétillante	☐	**2**	Plutôt légère en alcool	☐	**2**
Un goût plutôt typé	☐	**3**	Une mousse abondante	☐	**3**
Un goût plutôt neutre	☐	**4**	Peu de mousse	☐	**4**
Plutôt amère	☐	**5**	Un goût fruité	☐	**5**
Plutôt peu amère	☐	**8**			
Une odeur forte	☐	**7**			
Une odeur discrète	☐	**8**			
Qui laisse un arrière-goût	☐	**9**			
persistant	☐				
Qui ne laisse pas d'arrière-goût	☐	**0**			

QUESTIONNAIRE DE TEST

INSTRUCTIONS

Merci d'être venu aujourd'hui nous aider dans cette étude de marché sur Les bières. Il y a 5 produits à goûter.
Les produits vont vous être servis un à la fois, dans des verres marqués du code produit.

1. Goûtez le produit qui va vous être servi.
2. Remplissez le questionnaire pour ce produit.
3. Faites une petite pause pour laisser votre palais se reposer.
4. Signalez à l'enquêteur que vous êtes prêt à recevoir le prochain produit.
5. Buvez un peu d'eau minérale et mangez un morceau de cracker pour vous rincer la bouche.
6. Recommencez au point 1 pour le nouveau produit qu'on vous apportera.

Ordre de présentation *1 2 3 4 5* **(6)**
(entourer l'ordre adéquat)

Code produit | | | | **(7)(8)**

Q1. Tout bien considéré, que pensez-vous globalement de cette bière ? Nous aimerions que vous donniez une note entre 0 et 10 selon votre opinion: 0 si vous avez une très mauvaise opinion, jusqu'à 10 si vous avez une très *bonne* opinion Utilisez les notes intermédiaires pour nuancer votre jugement

Donnez votre impression globale : **(9)**
(COCHEZ UNE SEULE CASE)

Je n'aime pas du tout ce produit — J'aime vraiment beaucoup ce produit

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
0 1 2 3 4 5 6 7 8 9 10

Q2- Pouvez-vous me dire ce qui vous plaît dans ce produit ?

…… **(10)**
…… **(11)**
…… **(12)**
…… **(13)**
…… **(14)**

Q3. Pouvez-vous me dire ce qui vous déplaît dans ce produit ?

…… **(15)**
…… **(16)**
…… **(17)**
…… **(18)**
…… **(19)**

Merci. Veuillez signaler à l'enquêteur que vous êtes prêt à recevoir Le prochain produit.

ANNEXE 4 : REFERENCES DES LOGICIELS

Logiciel	Ordre de prix	Distribution
ALCESTE pour Macintosh Analyse des Lexèmes Cooccurrents dans les Enoncés Simples d'un Texte.	Non communiqué	IMAGE (société privée) 55, avenue Louis Bréguet Bât. 7 31 400 Toulouse Tél. 05 61 54 61 00 Fax. 05 61 80 53 03
ALETH-IP	Non communiqué	GSI-ERLI (Société privée) Eric GAUSSIER 1, place des marseillais 94 227 Charenton Le Pont Cedex Tél. 01 48 93 81 21 e-mail : Eric.Gaussier@erli.gsi.fr
Alta Vista / Life topics	Serveur : 70 000 F HT Client : 200 F HT	Digital Equipment France 9/13, avenue du Lac B.P. 235 91007 Evry Cedex Tél: 01 69 87 51 11 fax: 01 69 87 13 60 Web. http://www.digital.com
CANDIDE LEXIMAPP	Non communiqué	TRANSVALOR Tél. 01 40 51 90 00
COATIS	Non communiqué	D. GARCIA EDF/DER IMA/TIEM 1, avenue du Général de Gaulle 92 141 CLAMART Tél. 01 47 65 37 44 e-mail. Daniela.Garcia@der.edfgdf.fr
DATAVIEW	20 000 F HT	CRRM Hervé ROSTAING Université d'Aix-Marseille III Faculté des sciences et techniques de St Jérôme 13397 MARSEILLE CEDEX 20 Tél. 04 91 28 87 46 Fax. 04 91 28 87 12 e-mail. Rostaing@crrm.univ-mrs.fr Web. http://crrm.univ-mrs.fr

ETHNOS	16 900 F HT module statistiques : 18 500 F HT	SOFT CONCEPT G. DANAGUEZIAN 71, cours Albert Thomas 69447 LYON Cedex 03 Tél. 04 78 54 64 00 Fax. 04 72 33 19 91 Web. http://www.soft-concept.com
FASTR (pour Macintosh et Unix systems, 1995)	Non commercialisé	C. JACQUEMIN (Universitaire) IRIN-IUT Nantes 3, rue du Maréchal Joffre 44 041 NANTES CEDEX 01 Tél. 02 40 30 60 52 / 02 49 61 49 85 Fax. 02 40 30 60 53 e-mail. jacquemin@iut-nantes.univ-nantes.fr
FU : Text Filter Utility pour PC (version 3.56, 1990)	Gratuit	David Lo (particulier) 4516, Albert Street BURNABY, B.C. CANADA V5C 2G5
HYPERBASE pour Macintosh (version 1.5, 1992)	800 F HT	Institut National de la Langue Française Etienne BRUNET (Universitaire) UFR Lettres 98, Bd Herriot 06 204 NICE Tél. 04 93 37 54 44 Fax. 04 93 37 54 45 e-mail. FRMOP11.BITNET
INFOTRANS pour PC (version 4.0, 1995)	5 980 F HT	Information & Communication Dietrich Rieth Alte Str. 66 D-79249 FREIBURG MERZHAUSEN Tél. 00 49 761 40 49 49 Fax. 00 49 761 45 90 730 e-mail. info.fr@Rieth.de
IOTA	Non communiqué	M.-F. BRUANDET Laboratoire Génie Informatique CLIPS-IMAG Université Joseph Fournier B.P. 53 38 041 Grenoble Cedex 9 Tél. 04 76 51 45 15 e-mail. Marie-France.Bruandet@imag.fr

LE SPHINX pour PC (version 2)	Primo : 5 800 F HT Plus : 7 550 F HT Lexica : 11 800 F HT Module de saisie : 500 F HT Module Internet : 1 450 F HT	Le Sphinx Développement (société privée) 13, chemin des Amarantes 74 600 SEYNOD Tél. 04 50 51 17 56 Fax. 04 50 51 86 18 e-mail. lesphinx@cyberaccess.fr Web. http://www.alma.fr/sphinx
LEXICO1 pour Macintosh (1994)	Non commercialisé	Laboratoire Lexicométrie & textes politiques de l'ENS de Fontenay-Saint-Cloud André SALEM Av. de la Grille d'Honneur Le parc 92 211 SAINT-CLOUD Cedex Tél. 01 47 71 91 11 e-mail. salem@alize.msh-paris.fr
LEXIMAPPE	Non commercialisé	Jean-Pierre Courtial Centre de sociologie de l'innovation École des Mines de Paris 60 Bd Saint-Michel 75272 Paris Cedex 06 Tél. 01 40 51 90 00 Fax 01 43 25 94 95
LEXIS pour PC (EOLE version 3, 1994)	14 500 F HT Module statistique : 24 500 F HT	EOLE (société privée) 3bis, rue des Galons 92 190 MEUDON Tél. 01 46 26 80 00 Fax. 01 46 23 93 71
MATRISME	Non communiqué	Centre de recherche LePont Eric BOUTIN Université de Toulon-Var BP 132 83957 La Garde cedex Tél. 04 94 14 22 16 Fax. 04 94 14 22 75
NEUROTEXT	3 950 F HT	GRIMMER Logiciels 34bis, rue de Dunkerque 75010 PARIS Tél. 01 42 80 93 37 Fax. 01 42 80 93 39 Compuserve : 100073.155
NOEMIC	Non communiqué	C. KRUMEICH Société MA.NOS 18, rue Albert Einstein 77 420 Champs-sur-Marne Tél. 01 64 61 70 07

PAPINS : Prototype d'Analyse pour la Production d'INdex Sémantiques	Non communiqué	Florence PUGEAULT (Universitaire) IRIT, Université Paul Sabatier 118, route de Narbonne 31 062 Toulouse Cedex Tél. 05 61 55 62 44 Fax. 05 61 55 62 58 e-mail. pugeault@lexique.irit.fr
PHRASEA pour Macintosh	Non communiqué	B&L PARENTHESES (Société privée) R. Monchet 79, Av. Gyunemer 59 700 Marcq en Barœul Tél. 04 20 06 22 22 Fax. 04 20 12 07 40
POLLUX pour PC	Poste principal : 25 000 F HT Poste enquêteur : 4 500 F HT	AXIOM Software 130, bd Camélinat 92240 MALAKOFF Tél. 01 40 84 84 04 Fax. 01 40 84 84 00 e-mail. Axiom@worldnet.net
QUESTION	Version de base : 3 950 F HT Version Pro : 8 590 F HT Module de question ouvertes : 3 500 F HT Module de tableaux de Bord : 3 950 F HT	GRIMMER Logiciels 34bis, rue de Dunkerque 75010 PARIS Tél. 01 42 80 93 37 Fax. 01 42 80 93 39 Compuserve : 100073.155
Search'97	Serveur : 70 000 F HT client : 200 F HT	VERITY Laurent Le Foll 14, place Marie Jeanne Bassot 92593 LEVALLOIS PERRET CEDEX Tél. 01 41 49 04 51 Fax. 01 40 89 09 81 e-mail. rep-fr@verity.com Web. http://www.verity.com
SDOC et NDOC	Non commercialisé	INIST - CNRS Xavier POLANKO et Luc GRIVEL 2, allée du Parc de Brabois 54514 Vandœuvre-lès-Nancy Cedex Tél. 03 83 50 46 00 Fax. 03 83 50 47 48 Web. http://www.inist.fr

SEEK	Non communiqué	C. JOUIS Laboratoire IDIST/CREDO Université Charles de Gaulle - Lille III B.P. 149 95 653 Villeneuve d'Ascq Cedex Tél. 01 20 41 62 30 e-mail. jouis@univ-lille3.fr
SERAPHIN (Station SUN, 1995)	Non communiqué	J. BERRY (Universitaire) CNRS/EHESS Université Paris I EDF/DER
SNR : Multi-string text Search'N'Replace pour PC(version 1.5, 1988)	10$	Graphics Unlimited Inc. Thomas A. LUNDIN (société privée) 3000 2nd Street N. Minneapolis, MN 55411 USA
SPAD. T pour PC (version 1.5, 1993) Système Portable pour l'Analyse des Données Textuelles	7 950 F HT	Centre Internationale Statistique Informatique Appliquées (société privée) 1, avenue Herbillon 94 160 Saint-Mandé Tél. 01 43 74 95 26 Fax. 01 43 74 17 29
SPIRIT SENSE pour PC en client/serveur	Non communiqué	Société T.GID groupe Technologies (Société privée) P. MORDINI 84-88, Bd de la Mission Marchand 92 411 COURBEVOIE Tél. 01 49 04 70 70 Fax. 01 43 33 94 23 e-mail. mordini@syfed-pa.refer.fr
SYLEX : SYLEX-BASE et SYLEX(L-LEX) pour PC et Système Unix	Module d'analyse syntaxique : 60 000 F HT Module complet : 100 000 F HT	LANGAGE NATUREL-INGENIA (Société privée) Frédéric PIGAMO Technopôle de Château Gombert Europarc bât. D 13 013 Marseille Tél. 04 91 11 70 10 Fax. 04 91 11 75 77 e-mail. Pigamo@ingenia.fr
TETRALOGIE	60 000 F HT	Laboratoire IRIT Système d'information généralisé 118 route de Narbonne 31062 Toulouse Cedex, France Tel : (33) 05.61.55.63.23 e-mail. Chrisme@irit.fr , zurfluh@irit.fr, dousset@irit.fr

TEXT NAVIGATOR	200 000 F HT	ECAM / GBIS - IBM (société privée) Marie-Hélène ANTONI 68-76 Quai de la Rapée 75 592 Paris Cedex 12 Tél. 01 40 01 54 68 Fax. 01 49 28 08 60 e-mail. antoni@vnet.ibm.com
TEWAT pour RISC System/6000	50 000 F HT	ECAM - IBM (société privée) C. HUOT 68-76 Quai de la Rapée 75 592 Paris Cedex 12 Tél. 01 40 01 57 11 Fax. 01 49 28 08 60 e-mail. huot@vnet.ibm.com
TRI-DEUX pour PC (version 2.2, 1995)	Gratuit	Philippe CIBOIS (Universitaire) UFR de sciences sociales Université Paris V 12, rue Cujas 75 005 PARIS Tél. 01 43 75 26 63
Umap	Non communiqué	TRIVIUM Camille GUERMONPREZ BROUARD 10, bd Sébastopol 75004 PARIS Tél. 01 44 78 64 29 Fax. 01 44 78 64 30 e-mail. trivium@trivium.fr Web. http://www.umap.com
Vite Lu et autres produits de Technociel pour Macintosh (1995)	Non communiqué	Technociel (Société privé, consultant) Nicolas GERMAIN 18, rue Jubin 69 100 VILLLEURBANNE FRANCE Tél. 04 72 43 91 17 Fax. 04 78 93 02 74

ANNEXE 5 : LOGICIELS NON EVALUES SUR NOS DONNEES

Produit	Discipline	Rôle	Description
NOEMIC	Recherche documentaire	Codage	NOEMIC est la nouvelle appellation de TAÏGA. Ce système est conçu sur le principe de l'indexation noémique elle-même basée sur la représentation des informations. Cette indexation est réalisée grâce à des macro-structures considérées comme des unités sémantiques (concepts retranscrits en algèbre modale). Ce système est adapté à l'analyse de grands corpus hétérogènes de documents textuels.
SEEK	Linguistique	Codage	Ce système offre une aide à la modélisation des connaissances à partir de textes. Il est indépendant du domaine de connaissance. Le but est de guider le concepteur qui doit construire un modèle sans connaître le domaine décrit par les textes. Le système fonctionne grâce à des marqueurs linguistiques et des règles d'exploration conceptuelles. Les relations sémantiques sont issues du modèle linguistique de la grammaire applicative et cognitive.
IOTA	Recherche documentaire	Codage	Il est conçu sans connaissance sémantique a priori, sans analyseur morphosyntaxique et avec un lemmatiseur simple. Le système fonctionne si les corpus sont homogènes. Il est considéré comme un excellent filtre pour extraire et organiser le vocabulaire du domaine d'une façon simple.
COATIS	Recherche documentaire	Codage	Le système permet de relier des actions (verbes) à des causalités (les mots qui lui sont associés). Il fonctionne sur des textes traités par un extracteur de terminologie (ici, LEXTER). Il est indépendant d'un domaine d'application. Il ne réalise pas d'analyse syntaxique. Il fournit des éléments pour faciliter la compréhension des textes. Il est utile pour construire une base de connaissance à partir de textes.

SPIRIT SENCE	Recherche documentaire Linguistique	Codage	Ce système propose : • une extraction dynamique des idées contenues dans le texte (mode interactif) • une création automatique de sommaire avec des tables thématiques, des listes de description, des annotations analytiques, des balises hypertextes, des résumés... • une assistance à l'interrogation. Pour cela, il s'appuie sur l'analyse rhétorique du texte (définition du genre : délibératif, démonstratif, judiciaire/définition des ambiguïtés), sur l'extraction des concepts (873 concepts de base selon la logique câblée universelle) et sur la construction de la sémantique du texte (expressions algébriques représentant les relations sémantiques entre les concepts).
Atelier KES	Recherche documentaire Linguistique	Codage	Le système s'appuie sur la combinaison de plusieurs techniques : des méthodes et outils linguistiques des méthodes statistiques, des systèmes experts, des réseaux de neurones, des bases de données, des hypertextes. KES est une boîte à outils qui entre dans le cadre du projet GRAAL. Il permet d'extraire un ensemble structuré de données à partir de textes bruts et de répondre aux différents besoins d'application.
L4U DIGOUT4U IM4U	Recherche documentaire Linguistique	Codage Traitement statistique Représentation graphique	C'est un dérivé de TAÏGA. Language For You est multilingue et quoique s'appuyant sur une analyse sémantique, il intègre une analyse synthaxique. L'objectif de ce logiciel est de filtrer, sans risque d'erreur ou d'oubli, les textes non redondants.
DR-Link	Recherche documentaire	Codage	Encore appelé Document Retrievial through Linguistic Knowledgs. Il permet de déterminer automatiquement si des événements rapportés dans un texte sont effectivement survenus ou sont susceptibles de survenir. Les noms propres, par exemple, sont archivés sous 40 rubriques avec 2 niveaux de hiérarchie, de sorte que les liens créés permettent de retrouver intelligemment une information correspondant exactement à la demande. Les questions sont posées en langage naturel et font appel à des idées, ce qui permet d'effectuer des recherches dans des bases non structurées et totalement hétérogènes.

Périclès	Recherche documentaire	Collecte Codage	Il coordonne 6 moteurs de recherche qui cherchent simultanément dans plusieurs sources hétérogènes. Les informations recueillies sont réparties en deux catégories : l'information pertinente et l'information intéressante.
Sémiomap	Recherche documentaire Représentation graphique	Collecte Codage Traitement statistique Représentation graphique	Il indexe l'ensemble des pages Web sur le monde et sur cette base, fournit une sorte de "carte sémantique" sous la forme de diagramme montrant les liens entre un événement, un mot ou les mots qui lui sont associés. En tapant le mot recherché, le logiciel fait donc apparaître sur l'écran une carte avec des vignettes de couleurs différentes chacune représentant un "agrégat" statistique de mots qui apparaissent régulièrement ensemble dans le même contexte. On a alors une vision synthétique du contenu des pages.
Gingo	Recherche documentaire Représentation graphique	Codage Traitement statistique Représentation graphique	C'est un logiciel de management et de cartographie des ressources humaines et des informations stratégiques des entreprises. Il fonctionne sur le principe des "arbres de connaissances".
Name Tag	Recherche documentaire	Codage Traitement statistique	Le logiciel est incorporé dans Netowl qui est un moteur de recherche sur le Web. Ce dernier est conçu pour chercher uniquement les noms propres parmi lesquels il distingue les personnalités, les sociétés, les lieux, les expressions monétaires, etc…Opérant à la vitesse de 35 000 caractères/seconde, il construit des index sur mesure, établit des liens hypertextes entre l'index et les fichiers et prépare des résumés.

TETRALOGIE	Recherche documentaire Représentation graphique	Collecte Codage Traitement statistique Représentation graphique	Tétralogie est un outil de découverte de connaissances cachées dans une masse de données. Un tel processus débute par la sélection de données (indicateurs stratégiques) dans différentes bases, éventuellement hétérogènes. Cette sélection peut faire intervenir des filtres : négatifs, ils permettent d'éliminer des éléments d'information inutiles; positifs, ils permettent de sélectionner des éléments pertinents. La sélection peut également faire intervenir des dictionnaires hiérarchisés qui résolvent les problèmes de synonymie, de spécificité et de généricité au niveau des valeurs des différentes caractéristiques retenues pour l'étude. Ces données sont ensuite analysées dans le but d'en extraire des informations endogènes (cachées), grâce à des mécanismes complémentaires de découverte de connaissance : × recherche de dépendances entre variables (co-occurrence, recoupement, analyse multidimensionnelles) par création de tables de croisements disjonctives ou de contingence, × classification Cette découverte de connaissances cachées repose sur des méthodes statistiques et des méthodes d'analyse de données : Analyse en Composantes Principales, Analyse en Composantes Principales Réduites, Analyse Factorielle des Correspondances,.... C'est un système ouvert qui offre la possibilité d'ajouter de nouvelles fonctions d'analyse simplement. L'analyse est complétée par un module de visualisation sous forme de cartes factorielles à 2, 3 ou 4 dimensions qui offre une puissance de représentation grâce à laquelle l'utilisateur peut participer activement à la découverte de connaissance.

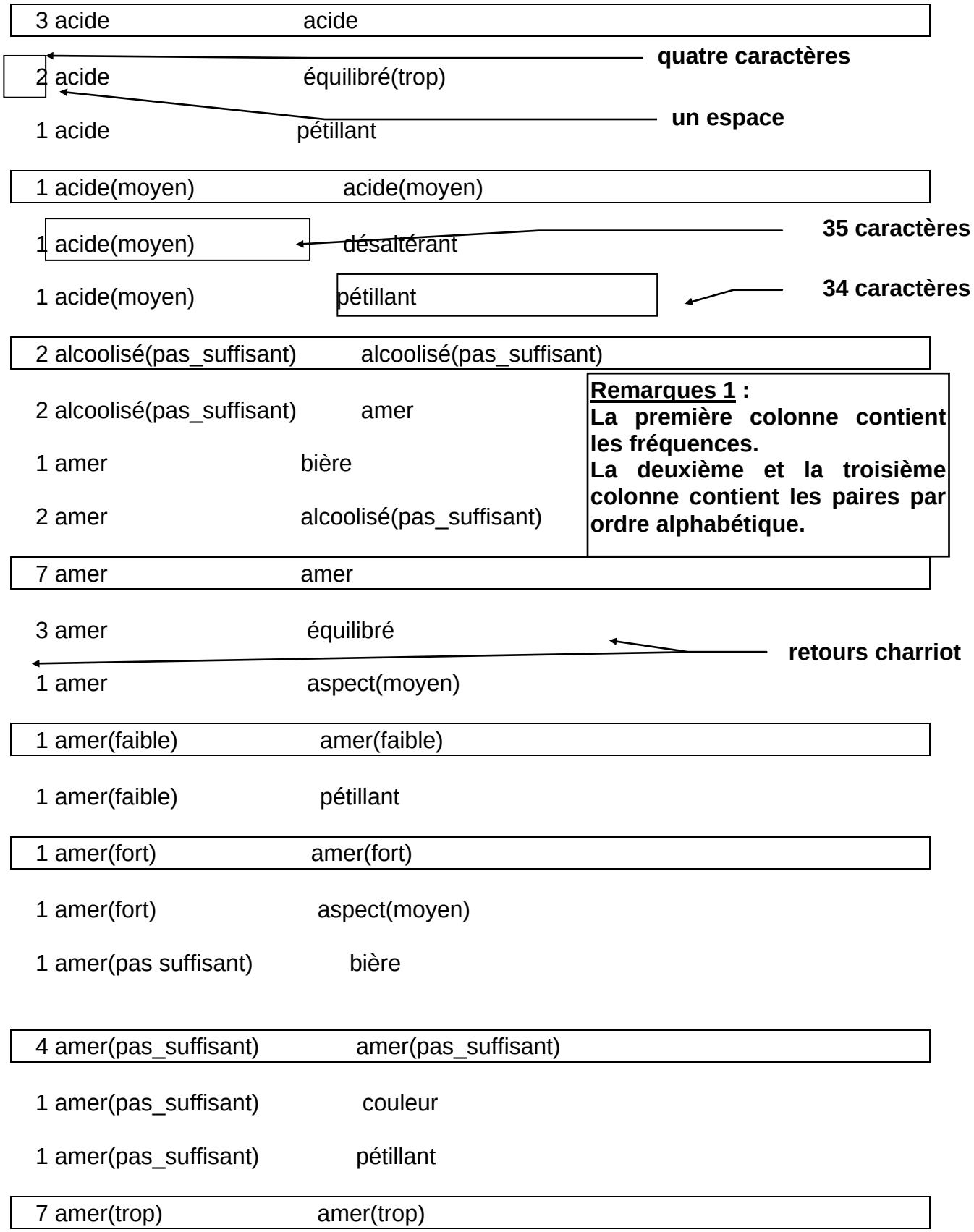
QUESTION	Traitement d'enquêtes	Collecte Codage Traitement statistique Représentation graphique	Version de base : × Paramétrage des questionnaires en questions simples, multiples, numériques, ouvertes. × Saisie des réponses avec contrôle des filtres, distinction entre les non-réponses et les non-concernés. × Codage des variables en classes d'intervalles, regroupements de modalités, sous-populations, création de sous-totaux. × Tris à plat, tris croisés présentés sous la forme de tableaux prêts à être insérés dans les rapports d'études, tests statistiques. × Graphiques en 2 ou 3 dimensions, secteurs, nuages de points... × Importations/Exportations de fichiers sous différents formats Version Pro : × **Macros** procédures pour les codages et les plans de tris. × **Lexicographie Fusion et extraction** : - horizontale : permet de regrouper les questions de plusieurs enquêtes. - verticale : permet de regrouper dans un seul fichier les questionnaires saisis sur plusieurs postes. × **Analyse de données** Analyses factorielles en composantes principales, des correspondances simples (ou binaires), des correspondances multiples. - Mappings en 2 ou 3 dimensions. - Typologie, classification hiérarchique ascendante, régression multiple, analyse factorielle discriminante, segmentation. × **Analyse conjointe** (trade-off Omix). × **Saisie multipostes**. × **Edition automatique des tableaux de bord**. Version avec Neurotex : × Lexicographie, statistique des mots, comptage, × Analyse de contenu, découpage, classification des idées, × Analyse des données textuelles (analyse factorielle, classification hiérarchique), × Identification des champs sémantiques à l'aide de réseaux de neurones.

LEXIS	Traitement d'enquêtes	Collecte Codage Traitement statistique Représentation graphique	Module disponible dans EOLE.3 pour analyser les questions ouvertes dans les enquêtes. Il est également utilisable pour l'étude de textes.
SAMPLER	Recherche documentaire	Collecte Codage Traitement statistique	SAMPLER™ est une boîte à outils d'analyse lexicométrique. Ce produit est indépendant de la langue et du domaine applicatif. Ses API facilitent une intégration de type OEM. Extraction terminologique × Calculs d'index, × Calculs de segments répétés, × Importation de lexiques, × Gestion de la substitution, × Navigation hypertextuelle, × Connexion à des outils d'analyse linguistique, × Extraction automatique de terminologie. Clustéristation paramétrable × Calculs de réseaux lexicaux, × Visualisation graphique interactive, × Export vers des couches de reformulation documentaire (Search'97™, Fulcrum,™, Spirit™, ...), × Export vers la plate-forme d'analyse et de suivi chronologique d'information développée par Cisi Utilisation : × Collecter l'information disponible : sites Internet, Intranet, forums "news groups", dépêches de presse, documentation de projets, rapports techniques, comptes rendus, ... × Etablir le lexique de référence en utilisant l'extracteur automatique de terminologie de SAMPLER™, × Analyser automatiquement ce "corpus" de textes avec SAMPLER™ en utilisant le lexique de référence, × Utiliser les réseaux lexicaux (clusters) obtenus pour naviguer et accéder intelligemment à la documentation via les moteurs de recherche standard du marché ou les exporter vers des environnements d'analyse spécialisés développés par Cisi (plate-forme de suivi chronologique de l'information , plate-forme d'analyse sémiotique).

U-MAP	Recherche documentaire Représentation graphique	Collecte Codage Traitement statistique Représentation graphique	Au cours d'une première étape, U-Map "aspire" de manière classique en utilisant plusieurs moteurs de recherche (Alta Vista, Yahoo, etc) les pages du Web correspondant aux mots clefs d'une recherche. Dans un deuxième temps, il indexe les mots contenus dans chaque page, et crée ainsi un corpus. En analysant celui-ci grâce à un algorithme qui s'appuie sur la proximité entre les mots (il n'a donc pas besoin de dictionnaire spécifique), il crée une cartographie des pages téléchargés.
ALETH	Recherche documentaire	Codage	Ce logiciel fait de l'indexation à plat (les index sont des chaînes de caractères).
SERAPHIN	Recherche documentaire	Codage	Ce logiciel extrait les phrases les plus significatives d'un texte en utilisant les marqueurs linguistiques laissés par l'auteur. Cette démarche est mise en œuvre à l'aide de la méthode d'exploration contextuelle développée au C.A.M.S. La méthode d'exploration contextuelle consiste d'une part, à repérer des indices linguistiques (lexèmes, marques grammaticales, de temps et marques structurelles) à l'intérieur de leur contexte textuel et d'autre part, à exprimer des règles heuristiques qui statuent sur la fonction et l'importance d'un énoncé, en s'appuyant sur ces indices linguistiques préalablement repérés.
FASTR	Recherche documentaire	Codage	FASTR est un logiciel d'extraction terminologique qui nécessite une ressource comme un thesaurus. Il met en œuvre des techniques syntaxiques et lexicographiques à partir des termes du thesaurus. Ainsi, il peut prendre en compte les termes synonymes du thesaurus
HYPERBASE	Lexicométrie	Codage Traitement statistique	Il produit un dictionnaire des formes graphiques interactif qui se présente selon un ordre alphabétique mais qu'il est possible de transformer en une liste des fréquences décroissantes. Ce logiciel repose sur une analyse comparative de la fréquence des formes graphiques. La première comparaison se rapporte à un corpus externe, un extrait du Trésor de la Langue Française, la seconde au corpus lui-même. Les premières statistiques produisent un tableau de la richesse lexicale de chaque entretien (norme du TLF). Hyperbase produit trois types de fichiers de vocabulaire spécifique. On peut effectuer, sous Hyperbase, des AFC sur des formes graphiques par l'intermédiaire du programme ADDAD. Elles portent sur des listes limitées, préétablies de différentes façons : par les fréquences, par la longueur des formes, à partir du vocabulaire spécifique, par thèmes, ...

PHRASEA	Recherche documentaire	Codage	Logiciel professionnel d'analyse et d'archivage multimédia en texte intégral. Archive tous documents textuels, graphiques, sonores ou animés et dispose d'outils puissants permettant de les retrouver et de les analyser : opérateurs logiques et géographiques, mots vides, dictionnaires, synonymes, glossaires hiérarchiques, détail des recherches, structuration des bases, plusieurs modes de visualisation ...
PAPINS	Analyse linguistique	Codage	Ce prototype a pour objectif de fournir une représentation structurée du contenue des textes décrivant les projets de recherche de la DER d'EDF. Cet outil expérimental d'extraction de connaissances à partir de texte s est basé sur des descriptions de la sémantique lexicale.
OMNIMARK	Recherche documentaire	Codage	Langage de programmation pour préparer la mise au format SGML des documents.

ANNEXE 6 : EXEMPLE DE FICHIER DE SORTIE TEXTO



4 amer(trop) couleur

2 amer(trop) équilibré

1 amer(trop) pétillant

1 aspect(moyen) amer

1 aspect(moyen) amer(fort)

Remarques 2 :
Les lignes encadrées correspondent aux formes et à leurs fréquences : la première colonne contient la fréquence , la seconde et la troisième contiennent la même forme.

2 aspect(moyen)	aspect(moyen)

2 bière	bière

1 bière amer

1 bière amer(pas suffisant)

1 boire(faible)	boire(faible)

1 boire(faible) sucré(trop)

2 caractère fade

2 caractère	caractère

1 caractère goût(faible)

1 couleur amer(pas_suffisant)

4 couleur amer(trop)

5 couleur	couleur

1 désaltérant acide(moyen)

11 désaltérant	désaltérant

1 désaltérant pétillant

2 fade	fade

1 fade caractère

1 fade goût(faible)

1 fade frais(pas_suffisant)

1 frais(pas_suffisant) fade

1 frais(pas_suffisant) frais(pas_suffisant)

1 goût(faible) fade

1 goût(faible) fade(moyen)

1 goût(faible) caractère

2 goût(faible) goût(faible)

1 pétillant acide

1 pétillant acide(moyen)

1 pétillant amer(faible)

1 pétillant amer(pas_suffisant)

1 pétillant amer(trop)

1 pétillant désaltérant

5 pétillant pétillant

2 raffiné(moyen) raffiné(moyen)

1 raffiné(moyen) équilibré

1 raffiné(moyen) sucré(moyen)

1 raffiné(moyen) sucré(pas_suffisant)

1 sucré(moyen) raffiné(moyen)

1 sucré(moyen) équilibré

1 sucré(moyen) sucré(moyen)

1 sucré(pas_suffisant) raffiné(moyen)

1 sucré(pas_suffisant) sucré(pas_suffisant)

1 sucré(trop) boire(faible)

1 sucré(trop) sucré(trop)

1 équilibré	raffiné(moyen)
3 équilibré	amer
2 équilibré	amer(trop)
6 équilibré	équilibré
1 équilibré	sucré(moyen)
2 équilibré(trop)	acide
2 équilibré(trop)	équilibré(trop)

ANNEXE 7 : LISTE DES TERMES AMBIGUS

Ambiguïté	Solution	Exemple
§vraiment	∅	• **Vraiment** trop léger.
	très	• La première gorgée est **vraiment** amère.
	réel	• Est-ce **vraiment** de la bière ?
§tendre	ressemblance	• Elle **tend** à ...
	onctueux	• La mousse est **tendre**.
§température	∅	• **Température** fraîche qui contre une température ambiante.
	température	• Se boit à **température** idéale.
§spécial	∅	• Pas de sensation **spéciale**.
	pas classique	• Goût **spécial**.
	spécial	• Ressemble à une bière **spéciale.**
	très	• Goût **spécialement** désagréable.
§soutenir	∅	• Je **soutiens** cette idée.
	intense	• Odeur ou couleur **soutenue**.
§sentir	∅	• ressentir, avoir la sensation de ...
	odeur	• humer, respirer.
§sensation	∅	• Une **sensation** d'amertume un peu faible.
	sensation	• Peu de **sensations**.
§rester	∅	• Le **reste** est une histoire de goût.
	persistant	• Idée de persistance.
	tenue	• La mousse ne **reste** pas.
	difficile	• **Reste** sur l'estomac.
§ressortir	∅	• Le côté houblonné **ressort**.
	tonique	• Manque de **ressort**.

§reposer	∅ reconstituant	• Poser à nouveau. • Elle est **reposante.**
§plutôt	∅ assez fort	• C'est **plutôt** un panaché. • Elle est **plutôt** amère.
§plus	∅ assez fort pas assez pas très moyen	• Sucre **plus** goût de citron. • Elle est **plus** foncée. • Pourrait avoir un fruité un peu **plus** développé. • **Plus** besoin de boire. • La **plus** désaltérante est l'eau. • **Plus** ou moins
§plat	plat repas faible	• Cette bière est trop **plate**. • Bière à boire avec un **plat**. • Odeur **plate**.
§petite	∅ faible bonmarché petite	• En **petite** réception. • Une **petite** amertume. • **Petite** bière. • Bulles trop **petites**.
§passer	∅ pas persistant boire périmé	• **Passe** bien en petite réception. • Amertume **passant** vite. • **Passe** bien. • Goût **passé**.
§paille	paille jaune faible	• Odeur ou goût de paille. • Couleur paille
§moyen	∅ moyen	• Se laisse boire sans **moyens**. Peu de **moyens** financiers. • Amertume **moyenne.**

§mousseux	vin mousse	• Goût du **mousseux.** • Aspect **mousseux.**
§moins	pas un peu trop assez faible moyen	• **Moins** d'arôme. • Pourrait être **moins** alcoolisée. • On dirait du panaché avec **moins** de sucre. • Plus ou **moins**.
§mieux	assez fort pas idéal satisfaisant	• Une boisson chaude apaise **mieux**. • Il y a **mieux**. • Passe **mieux**.
§même	∅ resssemblance	• **même** si .... • **même** goût que ...
§meilleur	satisfaisant idéal	• **meilleur** goût. • C'est la **meilleur**.
§lourd	dense fort difficile intense	• Contraire de léger. • Mousse **lourde**. • **Lourd** à digérer. • Odeur **lourde**.
§longtemps	∅ persistant fort	• **Longtemps** après, ... • Une amertume en bouche plus **longtemps**. • Tient **longtemps** dans le verre.
§long	∅ persistant	• Ecœurant à la **longue**. • **Long** en bouche.
§léger	léger onctueux faible	• **Légèrete**. • Mousse **légère**. • **Légère** amertume.

§laisser	∅ pas satisfaisant persistant facile	• Me **laisse** indifférent. • **Laisse** à désirer. • **Laisse** un arrière-goût. • Elle se **laisse** boire.
§juste	∅ pas pas suffisant idéal	• Trop classique pour être bue **juste** pour le plaisir. • **Juste** assez d'amertume. • Qualité un peu **juste**. • **Juste** comme il faut. **Juste** assez. **Juste** bien.
§insipide	goût pas	• Bière **insipide**. • Couleur **insipide**.
§gros	∅ fort pas petite	• Pas de **gros** défauts. • Une **grosse** mousse. • **Grosses** bulles.
§goût	∅ satisfaisant goût	• **Goût** trop neutre. Trop amère, à mon **goût**. • A mon **goût**. • Manque de **goût**.
§fruité	∅ fruité	• Porter ses **fruit,** réussir. • Arôme de **fruit**.
§forte	intense fort	• Assez **forte**. • Amertume **forte.**
§force	pas plaisant intense fort	• Je me suis **forcé** à la boire. • Elle possède une certaine **force**. • **Force** en goût.

§fin	∅ raffiné arrièregoût petite mauvais	• Vers la **fin**. • Amertume **fine**. • En **fin** de bouche. • **Fines** bulles. • C'est la dernière de toutes...
§fade	fade faible	• Saveur. • Couleur **fade**.
§été	∅ été	• Le verbe être. • La saison.
§dur	pas facile persistant âpre	• **Dur** à avaler. • La mousse **dure** assez longtemps. • Goût **dur**.
§doux	doux faible	• Saveur sucrée. • Couleur **douce**.
§descendre	∅ pas tenue boire facile	• On la sent **descendre.** • La mousse **descend** vite. • Elle **descend** bien.
§dense	fort dense	• Mousse pas très **dense**. • Manque de **densité**.
§dégustation	∅ dégustation	• Première **dégustation**. • Agréable à **déguster**.
§couper	∅ aqueux	• Toute signification autre de coupe-soif. • Elle semble avoir été **coupée**.
§chaud	agréable chaud	• Couleur **chaude**. • Contraire de froid.

§caractère	∅ caractère	• **Caractère** amère. • Manque de **caractère**.
§café	café bar	• odeur ou arôme. • Je la consomme dans un **café**.
§brûlant	brûlé agressif	• Arôme **brûlé.** • Texture **brûlante.**
§bouche	∅ bouche	• Drôle de goût en **bouche**. • Agréable en **bouche**.
§boire	∅ arrièregoût boire choisir faible satisfaisant moyen	• Pas l'impression de **boire** une bière... • Après l'avoir **boire**. • Facile à **boire**. • S'il n'y a rien d'autre à **boire**. • Buvable.
§bière	∅ bière	• **Bière** agréable à boire. • **Bière** assez prononcée.
§bien	∅ assez fort bienfaisant satisfaisant équilibré	• N'attire pas **bien** l'oeil. • Attire **bien** l'oeil. • Elle fait du **bien.** • Elle passe **bien**. • **Bien** fondant. **Bien** dosé.
§argent	argenté finance	• Couleur. • Si j'ai de **l'argent**.
§aspect	∅ aspect	• Regarder la télévision. • Bel **aspect**, **aspect** gazeux...

ANNEXE 8 : ECHELLE DE QUANTIFICATION (7 ET 3 NIVEAUX)

Terme d'origine	Echelle à 7 niveaux	Echelle à 3 niveaux
abondant	très_fort	fort
absolument	très_fort	fort
assez fort fort	très_fort	fort
assez fort grand	très_fort	fort
assez fort prononcé	très_fort	fort
aucun pas	très_fort	fort
beaucoup assez fort	très_fort	fort
de plus en plus	très_fort	fort
dominant	très_fort	fort
énorme	très_fort	fort
exagérer	très_fort	fort
excessivement assez fort fort	très_fort	fort
exclusif	très_fort	fort
extrême	très_fort	fort
extrêmement	très_fort	fort
fort prononcé	très_fort	fort
fortement prononcé	très_fort	fort
fourni	très_fort	fort
généreux	très_fort	fort
incroyable	très_fort	fort
large gamme	très_fort	fort
légèrement sursaturer	très_fort	fort
multitude	très_fort	fort
nettement	très_fort	fort
pas pas	très_fort	fort
plein de	très_fort	fort
prédominant	très_fort	fort
sans pas	très_fort	fort
saturer	très_fort	fort
sursaturer légèrement	très_fort	fort
tellement	très_fort	fort
total	très_fort	fort
très	très_fort	fort
très ample	très_fort	fort
très assez fort	très_fort	fort
très fort	très_fort	fort
très fortement	très_fort	fort
très grand	très_fort	fort
très marqué	très_fort	fort
très prononcé	très_fort	fort
très très	très_fort	fort
très très fort	très_fort	fort
un sacré	très_fort	fort
vraiment très	très_fort	fort
accentué	fort	fort

beaucoup	fort	fort
bien marqué	fort	fort
élevé	fort	fort
fort	fort	fort
fortement	fort	fort
franchement	fort	fort
grand	fort	fort
largement	fort	fort
légèrement assez fort	fort	fort
pas mal	fort	fort
peu pas	fort	fort
profond	fort	fort
prononcé	fort	fort
puissant	fort	fort
riche	fort	fort
sans modéré	fort	fort
assez	assez_fort	fort
assez bien prononcé	assez_fort	fort
assez fort	assez_fort	fort
assez marqué	assez_fort	fort
assez prononcé	assez_fort	fort
assez puissant	assez_fort	fort
légèrement prononcé	assez_fort	fort
soutenu	assez_fort	fort
non masqué	assez_fort	moyen
un peu élévé	assez_fort	moyen
un peu fort	assez_fort	moyen
un peu marqué	assez_fort	moyen
un peu prononcé	assez_fort	moyen
bof	moyen	moyen
égal	moyen	moyen
médiocre	moyen	moyen
modéré	moyen	moyen
modeste	moyen	moyen
moyen	moyen	moyen
moyennement	moyen	moyen
moyennement prononcé	moyen	moyen
relativement	moyen	moyen
assez faible	assez_faible	moyen
assez faible prononcé	assez_faible	moyen
assez moyen	assez_faible	moyen
assez pas	assez_faible	moyen
assez peu	assez_faible	moyen
assez peu prononcé	assez_faible	moyen
pas beaucoup	assez_faible	moyen
pas dominant	assez_faible	moyen
pas énormement	assez_faible	moyen
pas fort	assez_faible	moyen
pas franchement	assez_faible	moyen
pas prédominant	assez_faible	moyen

pas réellement	assez_faible	moyen
pas sans	assez_faible	moyen
pas spécialement prononcé	assez_faible	moyen
pas tellement	assez_faible	moyen
pas vraiment	assez_faible	moyen
sensiblement	assez_faible	moyen
un peu	assez_faible	moyen
un peu assez faible	assez_faible	moyen
un peu assez fort	assez_faible	moyen
assez discret	assez_faible	faible
pas fortement	assez_faible	faible
pas très	assez_faible	faible
pas très fort	assez_faible	faible
pas très marqué	assez_faible	faible
pas très prononcé	assez_faible	faible
petit	assez_faible	faible
pratiquement peu	assez_faible	faible
relativement faible	assez_faible	faible
relativement peu	assez_faible	faible
très moyen	assez_faible	faible
un petit peu	assez_faible	faible
un peu faible	assez_faible	faible
bref	faible	faible
chouia	faible	faible
dépens	faible	faible
discret	faible	faible
exceptionnel	faible	faible
faible	faible	faible
faiblement	faible	faible
feutré	faible	faible
inhibé	faible	faible
légèrement	faible	faible
limité	faible	faible
nettement assez faible	faible	faible
pas grand	faible	faible
pas un grand	faible	faible
peu	faible	faible
peu accentué	faible	faible
peu chargé	faible	faible
peu élevé	faible	faible
peu fort	faible	faible
peu marqué	faible	faible
peu prononcé	faible	faible
pointe	faible	faible
sans être	faible	faible
sans être prononcé	faible	faible
sans grand	faible	faible
superficiel	faible	faible
tantinet	faible	faible
très discret	faible	faible

très faible	faible	faible
très légèrement	faible	faible
très petit	faible	faible
très peu	faible	faible
très peu marqué	faible	faible
très très légèrement	faible	faible
très très peu	faible	faible
un brin	faible	faible
un petit pas	faible	faible
zeste	faible	faible
à peine	très_faible	faible
à peine prononcé	très_faible	faible
absent	très_faible	faible
absent total	très_faible	faible
amoindrir	très_faible	faible
assez fort faible	très_faible	faible
assez fort pas	très_faible	faible
assez fort peu	très_faible	faible
aucun	très_faible	faible
dépourvu	très_faible	faible
en rien	très_faible	faible
franchement pas	très_faible	faible
jamais	très_faible	faible
minimum	très_faible	faible
non	très_faible	faible
non pas	très_faible	faible
pas	très_faible	faible
pas du tout	très_faible	faible
pas marqué	très_faible	faible
pas prononcé	très_faible	faible
pas puissant	très_faible	faible
pas très pas	très_faible	faible
plus du tout	très_faible	faible
pratiquement pas	très_faible	faible
rien	très_faible	faible
sans	très_faible	faible
sans aucun	très_faible	faible
sans dominant	très_faible	faible
sans force	très_faible	faible
sans pétulance	très_faible	faible
sans que	très_faible	faible
sans rien	très_faible	faible
sauf	très_faible	faible
très pas	très_faible	faible
très très pas	très_faible	faible
très très très légèrement	très_faible	faible
un faible petit	très_faible	faible
un peu pas	très_faible	faible
un tout petit léger	très_faible	faible
vraiment pas	très_faible	faible

ANNEXE 9 : ECHELLE DE JUGEMENT (5 ET 3 NIVEAUX)

Terme d'origine	Echelle à 7 niveaux	Echelle à 3 niveaux
beaucoup trop	trop	trop
beaucoup trop prononcé	trop	trop
excès	trop	trop
excessif	trop	trop
excessivement	trop	trop
manque pas	trop	trop
trop fort	trop	trop
trop grand	trop	trop
trop important	trop	trop
trop	trop	trop
trop marqué	trop	trop
trop prononcé	trop	trop
vraiment trop	trop	trop
légèrement trop	un_peu_trop	trop
un peu trop	un_peu_trop	trop
un peu trop fort	un_peu_trop	trop
un peu trop marqué	un_peu_trop	trop
un peu trop prononcé	un_peu_trop	trop
un petit peu trop	un_peu_trop	suffisant
un petit peu trop fort	un_peu_trop	suffisant
un tout petit peu trop	un_peu_trop	suffisant
complet	suffisant	suffisant
pas excès	suffisant	suffisant
pas excessif	suffisant	suffisant
pas trop	suffisant	suffisant
pas trop discret	suffisant	suffisant
pas trop grand	suffisant	suffisant
pas trop fort	suffisant	suffisant
pas trop marqué	suffisant	suffisant
pas trop pas	suffisant	suffisant
pas trop peu	suffisant	suffisant
pas trop prononcé	suffisant	suffisant
pas trop puissant	suffisant	suffisant
sans être excessif	suffisant	suffisant
sans être trop	suffisant	suffisant
sans être trop fort	suffisant	suffisant
sans excès	suffisant	suffisant
sans trop	suffisant	suffisant
sans un trop fort	suffisant	suffisant
suffisant	suffisant	suffisant
à peine assez	un_peu_insuffisant	pas_suffisant
faible manque	un_peu_insuffisant	pas_suffisant
leger manque	un_peu_insuffisant	pas_suffisant
légèrement pas assez	un_peu_insuffisant	pas_suffisant
légèrement trop faible	un_peu_insuffisant	pas_suffisant

manquant un léger	un_peu_insuffisant	pas_suffisant
manquant un peu	un_peu_insuffisant	pas_suffisant
manque légèrement	un_peu_insuffisant	pas_suffisant
manque un peu	un_peu_insuffisant	pas_suffisant
manque un pointe	un_peu_insuffisant	pas_suffisant
manque un tout petit peu	un_peu_insuffisant	pas_suffisant
pas assez	un_peu_insuffisant	pas_suffisant
pas assez fort	un_peu_insuffisant	pas_suffisant
pas assez marqué	un_peu_insuffisant	pas_suffisant
pas assez prononcé	un_peu_insuffisant	pas_suffisant
pas excessivement	un_peu_insuffisant	pas_suffisant
pas suffisamment	un_peu_insuffisant	pas_suffisant
pas suffisant	un_peu_insuffisant	pas_suffisant
sans beaucoup	un_peu_insuffisant	pas_suffisant
sans en être	un_peu_insuffisant	pas_suffisant
trop moyen	un_peu_insuffisant	pas_suffisant
un petit manque	un_peu_insuffisant	pas_suffisant
un peu insuffisant	un_peu_insuffisant	pas_suffisant
un peu pas assez	un_peu_insuffisant	pas_suffisant
un peu trop faible	un_peu_insuffisant	pas_suffisant
insuffisant	trop_peu	pas_suffisant
manque	trop_peu	pas_suffisant
manque de franchise	trop_peu	pas_suffisant
manque nettement	trop_peu	pas_suffisant
manque total	trop_peu	pas_suffisant
manque un peu pas	trop_peu	pas_suffisant
pauvre	trop_peu	pas_suffisant
sur sa faim	trop_peu	pas_suffisant
très insuffisant	trop_peu	pas_suffisant
trop discret	trop_peu	pas_suffisant
trop faible	trop_peu	pas_suffisant
trop pas	trop_peu	pas_suffisant
trop petit	trop_peu	pas_suffisant
trop peu	trop_peu	pas_suffisant
un peu trop pas	trop_peu	pas_suffisant

ANNEXE 10 : CLASSES DES TERMES DESCRIPTIFS EMPLOYES DANS LES COMMENTAIRES LIBRES

*Arôme*_____

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" YAOURT "	" @YAOURT "	" NOYAU "	" @NOYAU "
" WHISKY "	" @WHISKY "	" NOIX "	" @NOIX "
" VINEUX "	" @VIN "	" NOISETTE "	" @NOISETTE "
" VINBLANC "	" @VIN_BLANC "	" MOISISSURE "	" @MOISI "
" VINAIGRE "	" @VINAIGRE "	" MOISI "	" @MOISI "
" VIN "	" @VIN "	" MIELLEUX "	" @MIEL "
" FRUITDUVERGER "	" @FRUIT_DU_VERGER "	" MIEL "	" @MIEL "
" VAISSELLE "	" @VAISSELLE "	" METALLIQUE "	" @METAL "
" TOURBE "	" @TOURBE "	" MEDICAMENTS "	" @MEDICAMENT "
" TISANE "	" @TISANE "	" MEDICAMENTEUX "	" @MEDICAMENT "
" TERREUX "	" @TERRE "	" MEDICAMENT "	" @MEDICAMENT "
" TERRE "	" @TERRE "	" MALTEES "	" @MALT "
" TABAC "	" @TABAC "	" MALTEE "	" @MALT "
" SODA "	" @SODA "	" MALTE "	" @MALT "
" SIROPDERABLE "	" @SIROP_D_ERABLE "	" MALT "	" @MALT "
" SIROP "	" @SIROP "	" MALABAR "	" @MALABAR "
" SEIGLE "	" @SEIGLE "	" LIMONADE "	" @LIMONADE "
" SAVONS "	" @SAVON "	" LIEGE "	" @LIEGE "
" SAVON "	" @SAVON "	" LEVURE "	" @LEVURE "
" SARRASIN "	" @SARRASIN "	" KIRCH "	" @KIRCH "
" ROSE "	" @ROSE "	" JAVEL "	" @JAVEL "
" RHUM "	" @RHUM "	" JASMIN "	" @JASMIN "
" RESINEUX "	" @RESINE "	" HOUBLONNEE "	" @HOUBLON "
" RESINE "	" @RESINE "	" HOUBLONNE "	" @HOUBLON "
" REGLISSE "	" @REGLISSE "	" HOUBLON "	" @HOUBLON "
" RANCE "	" @RANCE "	" HERBE "	" @HERBE "
" RAISIN "	" @RAISIN "	" GROSEILLE "	" @GROSEILLE "
" POURRIT "	" @POURRI "	" GOUTMAIS "	" @MAIS "
" POURRIES "	" @POURRI "	" GIROFLE "	" @GIROFLE "
" POURRI "	" @POURRI "	" GINGEMBRE "	" @GINGEMBRE "
" POMMEVERTE "	" @POMME_VERTE "	" FUME "	" @FUME "
" POMMES "	" @POMME "	" FRUITS "	" @FRUITE "
" POMME "	" @POMME "	" FRUITROUGE "	" @FRUIT_ROUGE "
" POIVREE "	" @POIVRE "	" FRUITIER "	" @FRUITE "
" POIVRE "	" @POIVRE "	" FRUITES "	" @FRUITE "
" POIRE "	" @POIRE "	" FRUITEES "	" @FRUITE "
" PLATRE "	" @PLATRE "	" FRUITEE "	" @FRUITE "
" PLASTIQUE "	" @PLASTIQUE "	" FRUITE "	" @FRUITE "
" PLANTES "	" @PLANTE "	" FRUIT "	" @FRUITE "
" PISSEDEVACHE "	" @PISSE_DE_VACHE "	" FRAMBOISE "	" @FRAMBOISE "
" PHARMACIE "	" @PHARMACIE "	" FOINS "	" @FOIN "
" PHARMACEUTIQUE "	" @PHARMACIE "	" FOIN "	" @FOIN "
" PECHE "	" @PECHE "	" FLEURS "	" @FLEUR "
" PANACHE "	" @PANACHE "	" FLEURI "	" @FLEUR "
" OXYDE "	" @OXYDE "	" FLEUR "	" @FLEUR "
" ORGE "	" @ORGE "	" FARINE "	" @FARINE "
" ORANGER "	" @ORANGE "	" EXOTIQUES "	" @EXOTIQUE "
" ORANGE "	" @ORANGE "	" EXOTIQUE "	" @EXOTIQUE "
" OASIS "	" @OASIS "	" EPICES "	" @EPICE "
" NOYAUX "	" @NOYAU "	" EPICEE "	" @EPICE "

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" EPICE "	" @EPICE "	" CARTON "	" @CARTON "
" EAUSAVONNEUSE "	" @SAVON "	" CARBONE "	" @CARBONE "
" EAUMINERALE "	" @EAU_MINERALE "	" CARAMELISE "	" @CARAMELISE "
" DETERGENT "	" @DETERGENT "	" CARAMEL "	" @CARAMELISE "
" DECOMPOSITION "	" @DECOMPOSITION "	" CAOUTCHOUC "	" @CAOUTCHOUC "
" CUIVRE "	" @CUIVRE "	" CANNELLE "	" @CANNELLE "
" CITRONNEE "	" @CITRONNE "	" CALVADOS "	" @CALVADOS "
" CITRONNE "	" @CITRONNE "	" CAFE "	" @CAFE "
" CITRON "	" @CITRONNE "	" BRULE "	" @BRULE "
" CIRE "	" @CIRE "	" BONBON "	" @BONBON "
" CIDREE "	" @CIDRE "	" BOISE "	" @BOISE "
" CIDRE "	" @CIDRE "	" BLE "	" @BLE "
" CHIMIQUE "	" @CHIMIQUE "	" BERGAMOTE "	" @BERGAMOTE "
" CHEWINGGUM "	" @CHEWING_GUM "	" BANANE "	" @BANANE "
" CHAMPIGNON "	" @CHAMPIGNON "	" ARTIFICIEL "	" @ARTIFICIEL "
" CHAMPENOISE "	" @CHAMPAGNE "	" ARTICHAUT "	" @ARTICHAUT "
" CHAMPAGNE "	" @CHAMPAGNE "	" APERITIF "	" @APERITIF "
" CERISES "	" @CERISE "	" ANANAS "	" @ANANAS "
" CERISE "	" @CERISE "	" AMANDE "	" @AMANDE "
" CEREALES "	" @CEREALE "	" AGRUMES "	" @AGRUME "
" CEREALE "	" @CEREALE "	" AGRUME "	" @AGRUME "
" CASSIS "	" @CASSIS "		

*Hédonique*_____

FORME ORIGINALE	FORME ATTRIBUEE	FORME ORIGINALE	FORME ATTRIBUEE
" SYMPATHIQUE "	" *SYMPATHIQUE "	" IMMONDE "	" *IMMONDE "
" SYMPA "	" *SYMPATHIQUE "	" IMBUVABLE "	" *IMBUVABLE "
" SUPER "	" *SUPER "	" IGNOBLE "	" *IGNOBLE "
" SAVOUREUX "	" *SAVOUREUX "	" IDEAL "	" *IDEAL "
" SAVOUREUSEMENT "	" *SAVOUREUX "	" HORRIBLE "	" *HORRIBLE "
" SAVOUREUSE "	" *SAVOUREUX "	" HEUREUX "	" *HEUREUX "
" SAVOURER "	" *SAVOUREUX "	" HARMONIEUX "	" *HARMONIEUX "
" SAVOURE "	" *SAVOUREUX "	" GOULEYANT "	" *GOULEYANT "
" SATISFAISANT "	" *SATISFAISANT "	" GENIAL "	" *GENIAL "
" REPOUSSANT "	" *REPOUSSANT "	" FRAIS "	" *FRAIS "
" RAFRAICHIT "	" *RAFRAICHISSANT "	" FRAICHEUR "	" *FRAIS "
" RAFRAICHISSEMENT "	" *RAFRAICHISSANT "	" FRAICHES "	" *FRAIS "
" RAFRAICHISSANTE "	" *RAFRAICHISSANT "	" FRAICHEMENT "	" *FRAIS "
" RAFRAICHISSANT "	" *RAFRAICHISSANT "	" FRAICHE "	" *FRAIS "
" RAFRAICHIR "	" *RAFRAICHISSANT "	" EXTRAORDINAIREMENT "	" *EXTRAORDINAIRE "
" PLAIT "	" *PLAISANT "	" EXTRAORDINAIRE "	" *EXTRAORDINAIRE "
" PLAISIR "	" *PLAISANT "	" EXTRA "	" *EXTRAORDINAIRE "
" PLAISANTE "	" *PLAISANT "	" EXQUISE "	" *EXQUIS "
" PLAISANT "	" *PLAISANT "	" EXQUIS "	" *EXQUIS "
" PARFAIT "	" *PARFAIT "	" EXECRABLE "	" *EXECRABLE "
" ONCTUEUX "	" *ONCTUEUX "	" EXCITE "	" *EXCITANT "
" MEDIOCRE "	" *MEDIOCRE "	" EXCITANT "	" *EXCITANT "
" MAUVAISE "	" *MAUVAIS "	" EXCELLENTE "	" *EXCELLENT "
" MAUVAIS "	" *MAUVAIS "	" EXCELLENT "	" *EXCELLENT "
" JOLIE "	" *JOLI "	" EPOUVANTABLE "	" *EPOUVANTABLE "
" JOLI "	" *JOLI "	" ENVIE "	" *ENVIE "
" INSUPPORTABLE "	" *INSUPPORTABLE "	" ECOEURANT "	" *ECOEURANT "
" INFECTE "	" *INFECT "	" DETESTABLE "	" *DETESTABLE "
" INFECT "	" *INFECT "	" DESALTERE "	" *DESALTERANT "

FORME ORIGINALE	FORME ATTRIBUEE	FORME ORIGINALE	FORME ATTRIBUEE
" INFAME "	" *INFAME "	" DESALTERANTE "	" *DESALTERANT "
" DESALTERANT "	" *DESALTERANT "	" BEL "	" *BEAU "
" DESAGREABLEMENT "	" *DESAGREABLE "	" ATTIRE "	" *ATTIRANT "
" DESAGREABLE "	" *DESAGREABLE "	" ATTIRANT "	" *ATTIRANT "
" DEPLU "	" *DEPLAISANT "	" ATROCE "	" *ATROCE "
" DEPLAIT "	" *DEPLAISANT "	" APPRECIER "	" *APPRECIABLE "
" DEPLAISANTE "	" *DEPLAISANT "	" APPRECIEE "	" *APPRECIABLE "
" DEPLAISANT "	" *DEPLAISANT "	" APPRECIE "	" *APPRECIABLE "
" DELICIEUX "	" *DELICIEUX "	" APPRECIATION "	" *APPRECIABLE "
" DELICIEUSEMENT "	" *DELICIEUX "	" APPRECIABLE "	" *APPRECIABLE "
" DELICIEUSE "	" *DELICIEUX "	" APPRECI "	" *APPRECIABLE "
" DEGUEULASSE "	" *DEGOUTANT "	" APPETISSANT "	" *APPETISSANT "
" DEGOUTANT "	" *DEGOUTANT "	" APAISANT "	" *APAISANT "
" DEGOUT "	" *DEGOUTANT "	" ALLECHANTE "	" *ALLECHANT "
" DECEVANTE "	" *DECEVANT "	" ALLECHANT "	" *ALLECHANT "
" DECEVANT "	" *DECEPTION "	" AIMER "	" *AIMER "
" DECEPTION "	" *DECEPTION "	" AIMENT "	" *AIMER "
" CHALEUREUX "	" *CHALEUREUX "	" AIME "	" *AIMER "
" BONNES "	" *BON "	" AGREABLES "	" *AGREABLE "
" BONNE "	" *BON "	" AGREABLEMENT "	" *AGREABLE "
" BON "	" *BON "	" AGREABLE "	" *AGREABLE "
" BIEN "	" *BIEN "	" AFFREUX "	" *AFFREUX "
" BEURK "	" *BEURK "	" AFFREUSE "	" *AFFREUX "
" BERK "	" *BEURK "	" ADORE "	" *ADORE "
" BELLE "	" *BEAU "	" ABOMINABLE "	" *ABOMINABLE "

Perception

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" VOLATILE "	" |VOLATILE "	" GORGEES "	" |GORGE "
" SUBTIL "	" |SUBTIL "	" GORGE "	" |GORGE "
" SUAVE "	" |SUAVE "	" GLISSE "	" |GLISSANT "
" SENTIR "	" |SENTEUR "	" GLISSANT "	" |GLISSANT "
" SENTEURS "	" |SENTEUR "	" FRISSONS "	" |FRISSON "
" SENTEUR "	" |SENTEUR "	" ESTOMAC "	" |ESTOMAC "
" SENT "	" |SENTEUR "	" DENTS "	" |DENTAIRE "
" SENSATIONS "	" |SENSATION "	" DENTISTE "	" |DENTAIRE "
" SENSATION "	" |SENSATION "	" DENTAIRE "	" |DENTAIRE "
" SALIVER "	" |SALIVER "	" DEGUSTER "	" |DEGUSTATION "
" RELEVEE "	" |RELEVE "	" DEGUSTEES "	" |DEGUSTATION "
" RELEVE "	" |RELEVE "	" DEGUSTE "	" |DEGUSTATION "
" PARFUMEE "	" |PARFUME "	" DEGUSTATION "	" |DEGUSTATION "
" PARFUME "	" |PARFUME "	" DEGLUTITION "	" |DEGLUTITION "
" PARFUM "	" |PARFUME "	" BOUQUET "	" |BOUQUET "
" PAPILLEGUSTATIVE "	" |PAPILLE_GUSTATIVE "	" BOUCHE "	" |BOUCHE "
" PAPILLE "	" |PAPILLE "	" AVALER "	" |AVALER "
" PALAIS "	" |PALAIS "	" AVALE "	" |AVALER "
" ODORAT "	" |ODEUR "	" ARRIEREGOUT "	" |ARRIERE_GOUT "
" ODORANTE "	" |ODEUR "	" AROMES "	" |AROMATISE "
" ODEURS "	" |ODEUR "	" AROME "	" |AROMATISE "
" ODEUR "	" |ODEUR "	" AROMATISER "	" |AROMATISE "
" NEZ "	" |NEZ "	" AROMATISEE "	" |AROMATISE "
" FOSSENASALE "	" |NEZ "	" AROMATISE "	" |AROMATISE "
" LANGUE "	" |LANGUE "	" AROMATIQUES "	" |AROMATISE "
" INODORE "	" |INODORE "	" AROMATIQUE "	" |AROMATISE "
" GUSTATIVES "	" |GUSTATIF "	" AROMATE "	" |AROMATISE "

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" GUSTATIVE "	" |GUSTATIF "	" APRESGOUT "	" |ARRIERE_GOUT "
" GOSIER "	" |GORGE "		

Saveur

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" TANIN "	" £TANIN "	" DOUX "	" £DOUX "
" SUCREES "	" £SUCRE "	" DOUCEUR "	" £DOUX "
" SUCREE "	" £SUCRE "	" DOUCEREUX "	" £DOUX "
" SUCRE "	" £SUCRE "	" DOUCEREUSE "	" £DOUX "
" SAVEURS "	" £SAVEUR "	" DOUCEMENT "	" £DOUX "
" SAVEUR "	" £SAVEUR "	" DOUCEATRE "	" £DOUX "
" SALIN "	" £SALE "	" DOUCE "	" £DOUX "
" SALEE "	" £SALE "	" CORSEE "	" £CORSE "
" SALE "	" £SALE "	" CORSE "	" £CORSE "
" INSIPIDE "	" £INSIPIDE "	" CORPS "	" £CORPS "
" GROSSIERE "	" £GROS "	" BITTER "	" £BITTER "
" GROSSIER "	" £GROS "	" ASTRINGENT "	" £ASTRINGENT "
" GROSSE "	" £GROS "	" AMERTUME "	" £AMER "
" FINESSE "	" £FIN "	" AMERE "	" £AMER "
" FINES "	" £FIN "	" AMER "	" £AMER "
" FINEMENT "	" £FIN "	" AIGRELET "	" £AIGRE "
" FINE "	" £FIN "	" AIGREDOUX "	" £AIGREDOUX "
" FIN "	" £FIN "	" ACIDULEE "	" £ACIDULE "
" FADEUR "	" £FADE "	" ACIDULE "	" £ACIDULE "
" FADE "	" £FADE "	" ACIDITE "	" £ACIDE "
" FADASSE "	" £FADE"	" ACIDE "	" £ACIDE "
" EDULCORANTS "	" £EDULCORANT "		

Texture

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" VISQUEUX "	" §VISQUEUX "	" MOELLEUX "	" §MOELLEUX "
" VENTEE "	" §VENTE "	" MOELLEUSE "	" §MOELLEUX "
" VELOUTE "	" §VELOUTE "	" METAL "	" §METAL "
" TROUBLE "	" §TROUBLE "	" LISSE "	" §LISSE "
" SIRUPEUX "	" §SIRUPEUX "	" LIQUOREUX "	" §LIQUOREUX "
" SECHERESSE "	" §SEC "	" LIQUOREUSE "	" §LIQUOREUX "
" SECHE "	" §SEC "	" LIQUIDE "	" §LIQUIDE "
" SEC "	" §SEC "	" LIMPIDE "	" §LIMPIDE "
" SAVONNEUX "	" §SAVONNEUX "	" LIGHT "	" §LIGHT "
" RUGUEUX "	" §RUGUEUX "	" LESSIVEUX "	" §LESSIVEUX "
" RUGOSITE "	" §RUGUEUX "	" LAVASSE "	" §LAVASSE "
" RONDEUR "	" §ROND "	" IRRITE "	" §IRRITANT "
" ROND "	" §ROND "	" IRRITANT "	" §IRRITANT "
" RECHE "	" §RAPE "	" GRAS "	" §GRAS "
" RAPEUX "	" §RAPE "	" GRANULES "	" §GRANULE "
" RAPEUSE "	" §RAPE "	" GAZEUX "	" §GAZEUX "
" RAPE "	" §RAPE "	" GAZEUSE "	" §GAZEUX "
" PLATITUDE "	" §PLAT "	" GAZEIFIEE "	" §GAZEUX "
" PLATE "	" §PLAT "	" GAZEIFIE "	" §GAZEUX "
" PLAT "	" §PLAT "	" GAZCARBONIQUE "	" §GAZEUX "
" PIQUE "	" §PIQUANT "	" GAZ "	" §GAZEUX "
" PIQUANTE "	" §PIQUANT "	" FLUIDE "	" §FLUIDE "

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" PIQUANT "	" §PIQUANT "	" FERRAILLE "	" §FERRAILLE "
" PICOTEMENTS "	" §PICOTANT "	" FERMENTER "	" §FERMENTE "
" PICOTEMENT "	" §PICOTANT "	" FERMENTEE "	" §FERMENTE "
" PICOTE "	" §PICOTANT "	" FERMENTE "	" §FERMENTE "
" PICOTANT "	" §PICOTANT "	" FERMENTATION "	" §FERMENTE "
" PETILLEMENT "	" §PETILLANT "	" FER "	" §FER "
" PETILLE "	" §PETILLANT "	" FARINEUSE "	" §FARINE "
" PETILLANTE "	" §PETILLANT "	" EVENTEE "	" §EVENTE "
" PETILLANT "	" §PETILLANT "	" EVENTE "	" §EVENTE "
" PETILLANCE "	" §PETILLANT "	" EPAISSE "	" §EPAISSEUR "
" PERRIER "	" §PERRIER "	" EPAIS "	" §EPAISSEUR "
" PATEUX "	" §PATEUX "	" EFFERVESCENT "	" §EFFERVESCENCE "
" PATEUSE "	" §PATEUX "	" EAUDESTEZ "	" §GAZ "
" ONCTUOSITE "	" §ONCTUOSITE "	" EAU "	" §EAU "
" ONCTUEUSE "	" §ONCTUOSITE "	" DILUEE "	" §DILUE "
" MOUSSEUX "	" §MOUSSEUX "	" DILUE "	" §DILUE "
" MOUSSEUSE "	" §MOUSSANT "	" DEPOT "	" §DEPOT "
" MOUSSE "	" §MOUSSANT "	" DENSE "	" §DENSE "
" MOUSSANT "	" §MOUSSANT "	" DEGAZEIFIEE "	" §DEGAZEE "
" DEGAZEE "	" §DEGAZEE "	" ALCOOLISEE "	" §ALCOOLISE "
" CREMEUX "	" §CREMEUX "	" ALCOOLISE "	" §ALCOOLISE "
" CONSISTANTE "	" §CONSISTANT "	" ALCOOLIQUE "	" §ALCOOLISE "
" CONSISTANT "	" §CONSISTANT "	" ALCOOL "	" §ALCOOLISE "
" CONSISTANCE "	" §CONSISTANT "	" AIGREUR "	" §AIGRE "
" BULLEUX "	" §BULLEUX "	" AIGRE "	" §AIGRE "
" BULLES "	" §BULLEUX "	" AGRESSIVITE "	" §AGRESSIF "
" BULLE "	" §BULLEUX "	" AGRESSIVE "	" §AGRESSIF "
" BRUTE "	" §BRUT "	" AGRESSIF "	" §AGRESSIF "
" BRUT "	" §BRUT "	" AGRESSE "	" §AGRESSIF "
" AQUEUX "	" §AQUEUX "	" AEREE "	" §AERE "
" AQUEUSE "	" §AQUEUX "	" ACRETE "	" §ACRE "
" APRETE "	" §APRETE "	" ACRE "	" §ACRE "
" APRE "	" §APRETE "	" ACCROCHEUSE "	" §ACCROCHEUSE "
" ALLEGEE "	" §ALLEGE "	" ACCROCHE "	" §ACCROCHEUSE "

Aspect

FORME ORIGINALE	FORME ATTRIBUE	FORME ORIGINALE	FORME ATTRIBUE
" TROUBLES "	" &TROUBLE "	" COLOREE "	" &COLOREE "
" TROUBLE "	" &TROUBLE "	" COLORANT "	" &COLOREE "
" TERNE "	" &TERNE "	" CLAIRE "	" &CLAIR "
" TEINTEE "	" &TEINTE "	" CLAIR "	" &CLAIR "
" TEINTE "	" &TEINTE "	" BRUNES "	" &BRUNE "
" ROUSSE "	" &ROUSSE "	" BRUNE "	" &BRUNE "
" OPAQUE "	" &OPAQUE "	" BLONDEUR "	" &BLONDE "
" LAITEUX "	" &LAITEUX "	" BLONDES "	" &BLONDE "
" JAUNECLAIR "	" &JAUNE_CLAIR "	" BLONDE "	" &BLONDE "
" JAUNE "	" &JAUNE "	" BLOND "	" &BLONDE "
" FONCEE "	" &FONCE "	" BLANCHES "	" &BLANCHE "
" FLOUE "	" &TROUBLE "	" BLANCHE "	" &BLANCHE "
" DOREE "	" &DORE "	" BLANC "	" &BLANCHE "
" COULEURCLAIRE "	" &COULEUR_CLAIRE "	" AMBREE "	" &AMBRE "
" COULEUR "	" &COLOREE "	" AMBRE "	" &AMBRE "

ANNEXE 11 : NORME DE SAISIE DES COMMENTAIRES LIBRES DE CONSOMMATEURS

Voici quelques consignes à respecter lors de la saisie des commentaires libres de consommateurs. Elles sont indispensables pour la réussite de la phase de traitement.

① La saisie doit s'effectuer en mode minuscule avec les lettres accentuées :

é, è, à, ê, ù, î, ô, û...

② Saisir la ponctuation évidente (lorsqu'elle est spécifiée) et sous-entendue (lorsqu'elle est inexistante).

Ex :
- *Ponctuation évidente* :
  Pas assez parfumée. Un peu fade. Pas assez pétillante. Pas de mousse.
- *Ponctuation sous-entendue* :
  Goût moyen avec légère amertume → Goût moyen, avec légère amertume.

③ Distribuer les termes de description de part et d'autre d'une conjonction de coordination.

Ex :
- Manque de caractère et de pétillant → Manque de caractère et manque de pétillant.

④ Saisir les chiffre en toutes lettres :

Ex :
- 6eme bouteille → sixième bouteille ...

⑤ Saisir les abréviations en mots complets :

Ex :
- 30 mn → trente minutes ...

⑥ Ne rien saisir lorsque le questionnaire n'est pas rempli.

⑦ Saisir les mêmes commentaires qu'à la question précédante (opération *copier/coller*) lorsqu'un questionnaire comporte :

- idem

- Même chose qu’à la première bouteille ...

⑧ Ne pas mettre de majuscule en début de phrase ainsi qu’à la suite d’un point.

Ex :
- bière juste bien. équilibre entre le sucre et l’acide.

⑨ Préciser de quoi il s’agit lorsque la réponse se limite à un terme de quantification.

Ex :
- aucun → aucun défaut (s’il s’agit d’une réponse faite à propos des défauts du produit) ...

N'oubliez pas d'effectuer une vérification orthographique sur l'ensemble du fichier à la fin de la saisie.

Merci de votre compréhension.