Université de Nice - Sophia Antipolis
UPRESA "Bases, Corpus et Langage"

JADT 1998

4èmes Journées internationales d'Analyse statistique des Données Textuelles

Textes réunis par Sylvie Mellet

Avec la collaboration de
É. Brunet, M. Juillard, L. Lebart, A. Salem

Ouvrage publié avec l'aide
du Conseil Général des Alpes-Maritimes
de l'Université de Nice – Sophia Antipolis

1998

Comité scientifique / Programme Committee

Roger Establet	*Univ. Aix-Marseille*
Monica Bécue	*Univ. Polytechnic of Catalunya*
Sergio Bolasco	*Univ. de Rome 'La Sapienza'*
Étienne Brunet	*Univ. Nice - Sophia Antipolis*
Tullio De Mauro	*Univ. de Rome 'La Sapienza'*
Annibale Elia	*Univ. de Salerno*
Dominique Labbé	*Univ. de Grenoble*
Ludovic Lebart (président)	*CNRS, ENST, Paris*
Alain Lelu	*Univ. Paris 8*
Sylvie Mellet	*CNRS, Nice*
Charles Muller	*Univ. de Strasbourg*
Max Reinert	*CNRS Toulouse*
Jacques Rouault	*CRISS Grenoble*
André Salem	*Univ. Paris III 'Sorbonne Nouvelle'*

Comité d'organisation / Organizing Committee

Étienne Brunet	*UNSA*
Jacques Hammerschmitt	*CNRS/UNSA*
Michel Juillard	*UNSA*
Xuan Luong	*UNSA*
Sylvie Mellet (coordinatrice)	*CNRS/UNSA*
Jean-Pierre Regourd	*UNSA*
André Salem	*Univ. Paris III 'Sorbonne Nouvelle'*

Colloque réalisé avec le soutien de :

- Université Nice - Sophia Antipolis
- U.F.R. Lettres, Arts et Sciences Humaines de l'UNSA
- Conseil Général des Alpes-Maritimes
- Ville de Nice
- Fondation Sophia Antipolis

UPRESA "Bases, Corpus et Langage", INaLF
Faculté des Lettres
98, bd. Edouard-Herriot, B.P. 209
F-06204 Nice Cedex 3
Fax : (33) (0)4 93 37 54 45

Mise en pages réalisée par Pierrette Uboldi, CNRS - INaLF, Nice

UTILISATION DES QUESTIONS OUVERTES DANS LES TESTS CONSOMMATEURS EN ANALYSE SENSORIELLE

Hélène Ziegelbaum, Michel Rogeaux
TEPRAL
68 route d'Oberhausbergen
67 037 Strasbourg cedex

Hervé Rostaing
CRRM
Faculté des sciences et techniques de St Jérôme
13 397 Marseille cedex 20

Résumé

Le comportement du consommateur vis à vis d'un produit alimentaire intéresse de plus en plus les industries agro-alimentaires. La collecte de l'information consommateur est réalisée grâce à des questionnaires de dégustation.

On a eu l'occasion de remarquer des biais avec l'emploi des questions fermées dans les tests sensoriels consommateurs. En effet, ces derniers peuvent être influencés par les réponses qui leur sont proposées. De cette façon 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, les nouveaux questionnaires comportent des questions ouvertes. Ainsi, 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'agro-alimentaire. 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 pour obtenir de l'information homogène 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.

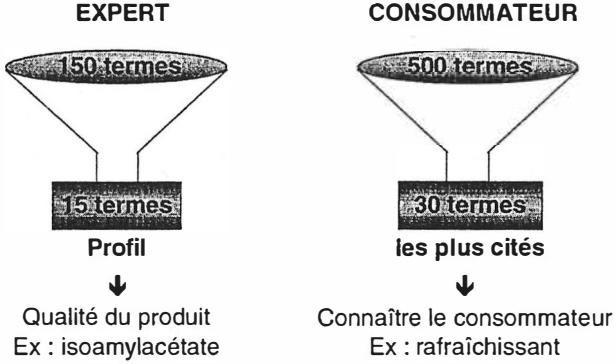
1. Contexte

Les contraintes de concurrence, d'exigence des consommateurs et de complexité de formulation obligent aujourd'hui les entreprises agro-alimentaires à s'engager dans trois types

d'activités : le développement de nouveaux produits, l'amélioration des produits existants et le maintien de la qualité de ces produits.

Mais le succès du produit dépend aussi du plaisir qu'il procure aux consommateurs et donc en grande partie de ses qualités organoleptiques.

Afin de comprendre pleinement et de maximiser au mieux l'acceptation d'un produit alimentaire au niveau sensoriel, deux types d'informations sont recherchées : celle des experts sensoriels et celle des consommateurs. Il s'agit dans les deux cas d'information textuelle. En effet, pour exprimer les sensations qu'ils ressentent après une dégustation de produit, expert et consommateur utilisent des critères sensoriels.



Le schéma ci-dessus montre une similarité dans la démarche chez l'expert et le consommateur. Par contre, le but et le mode de travail sont différents :

- avec les experts, on recherchera une adéquation entre la caractéristique sensorielle du produit et le descripteur ; il s'exprimera avec un vocabulaire spécifique.
- avec le consommateur, on recherchera une adéquation entre l'image du produit et son vocabulaire ; il s'exprimera avec son propre vocabulaire.

Les commentaires de dégustation constituent une source d'information assez riche et d'autant plus importante que les commentaires de consommateurs n'apparaissent en rien corrélés aux descriptions des experts (MARTIN N., ROGEAUX M., 1994). Le jury d'experts et le panel de consommateurs construisent deux représentations tout à fait indépendantes du produit. Les deux sont donc également stratégiques pour l'entreprise, le premier étant plus tourné vers le contrôle de la qualité industrielle et le second vers la «veille de marché ».

2. Méthode de traitement des commentaires libres

Elle se décompose en quatre phases :

- **acquisition** des données
- **postcodage** des données
- stockage et création d'index en **base de données**
- **calcul** et **représentation**

2.1 Acquisition

Dans l'optique de laisser parler librement le consommateur et de lui rendre sa spontanéité, nous utilisons une technique principalement employée dans les enquêtes socio-économiques où sont introduites des questions ouvertes. Les réponses du consommateur ne sont plus restreintes à une grille de choix limités mais se présentent sous forme de commentaires libres sur les produits.

La question se présente d'une manière simple, du type : «Vous venez de boire la bière X, quelles sensations vous procure-t-elle ?». 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.

2.2 Spécificité du poscodage

La stratégie de recherche sur le traitement des commentaires libres de consommateurs a été choisie en fonction d'un objectif bien défini qui est de conserver le minimum de termes pour un maximum de signification.

L'étape que nous.appelons postcodage va nous permettre d'accéder à cet objectif en proposant une « norme » de dépouillement des commentaires libres de consommateurs (LABBE D., 1990). Elle est basée sur une analyse morphologique classique comprenant les phases classiques de

- lemmatisation
- regroupement des locutions
- élimination des mots-vides
- regroupement synonymique
- levée d'ambiguïté

Ceci constitue une opération délicate qui consiste à reconnaître les formes et à constituer des dictionnaires de lemmes, de synonymes, de locutions, d'ambiguïtés et de mots vides.

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 (Information & Communication, 1995). Il s'agit d'un logiciel de reformatage de références bibliographiques. 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.

2.2.1. Lemmatisation

En lexicométrie, elle consiste à rattacher chaque mot du texte à une forme canonique et à une catégorie grammaticale. Dans notre cas, nous nous limiterons au regroupement des occurrences du texte sous une forme unique et stable. Le lexique est lemmatisé en prenant comme ordres de priorité :

- ramener les formes fléchies à l'adjectif ou aux participes masculin/singulier,
- ramener les formes fléchies au nom masculin/singulier,
- ramener les formes fléchies au verbe à l'infinitif,
- ramener les formes fléchies à la forme canonique.

Exemple :

Forme fléchie	Lemmes
astringents	astringent
astringentes	astringent
astringente	astringent
astringence	astringent

2.2.2. Regroupement synonymique

Une fois lemmatisés certains groupes de mots ont une signification voisine. Ils sont donc regroupés sous une seule dénomination.
Exemple :

Forme fléchie lemmatisée	Synonyme
rugueux	âpre
rude	âpre
râpeux	âpre
raide	âpre

2.2.3. Regroupement des locutions

Dans le langage usuel, un petit nombre de locutions est utilisé. Leur sens est important, il est donc nécessaire de les considérer comme une seule forme.
Dans le cas du postcodage, le groupe de mot sera relié par le caractère de soulignage « _ ».

Exemple :

Groupe de mots	Locution
tenue de mousse	tenue_de_mousse
bière sans alcool	bière_sans_alcool

2.2.4. Levée d'ambiguïté

L'ambiguïté lexicale provoque une incapacité d'identifier clairement le concept désigné par un mot surtout lorsqu'on travaille 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 on peut prévoir les cas classiques de polysémie et d'effectuer une analyse grammaticale intellectuelle.

Exemple :

Forme fléchie ambiguë	Significations
sentir	odeur ou sensation
plat	sans bulle ou un met
été	saison ou verbe être
doux	sucré ou faible

2.3. Elimination des mots vides :

Dans une phrase, certains mots sont plus chargés de sens sur le plan syntaxique que d'autres. On appelle mots vides ou encore mots outils, les termes de liaisons, les articles ou les éléments secondaires dans une phrase.

Ces derniers sont éliminés dans le même souci de concentration de l'information pertinente.

Exemple : à, car, cette, de, vers, tout, quelle, quand, puis, pendant, ...

2.4. Création d'index

Les données postcodées sont importées dans un logiciel documentaire TEXTO (Chemdata, 1996). Ce dernier permet non seulement de stocker les commentaires bruts et postcodés mais aussi de faire des calculs de paires entre deux mots appartenant au même commentaire. Cette opération est appelée création d'index dans TEXTO.

Les index contiennent l'ensemble des paires de termes avec leur fréquence d'apparition qui existe pour chaque profil.

2.5. Calcul et représentation

Cette application a été spécialement développée pour des besoins spécifiques :

- mettre en évidence les termes les plus fréquents
- mettre en évidence les liaisons les plus fortes entre deux termes
- utiliser la représentation en réseau de mots associés
- obtenir une représentation synthétique
- comparer un test à un autre
- en déduire une interprétation rapide et simple
- automatiser la construction.

Un certain nombre de choix ont donc du être faits lors de l'analyse des besoins notamment au niveau des indices, des associations et de la représentation.

2.5.1. Choix des types d'associations

Lors du calcul d'association des paires, on doit déterminer l'ensemble des mots sur lequel on choisi de calculer les associations.
M. REINERT, 1986 considère qu'une réponse ou un commentaire est une Unité de Contexte lui-même composé de plusieurs Unités de Contexte Elementaires ou UCE.
Pour G. TEIL, 1994, il s'agit d'unités de sens.
Si on veut identifier les principaux thèmes qui se dégagent dans le discours des consommateurs pour décrire un produit, on choisira de calculer les cooccurrences à partir des UCE.
Par contre, si on veut identifier les descripteurs qui sont associés dans le discours des consommateurs (par exemple, l'amertume est associée au désaltérant et au goût), on choisira de calculer les cooccurrences dans le commentaire entier.

2.5.2. Choix sur les indices

Sur une dizaines d'indices testés et comparés quatre ont été retenus pour les raisons suivantes

INDICE	INFORMATION MISE EN EVIDENCE
Indice de Jaccard	il favorise l'apparition des paires présentant une forte intensité de lien avec des fréquences d'apparition relativement faibles.
Indice de Russel & Rao	il met en évidence la fréquence relative de la paire.
Coéfficient de corrélation	il met en évidence deux types d'information : en positif les mots qui apparaissent toujours ensemble et en négatif les mots qui apparaissent toujours seuls.
Inclusion	il favorise les fortes fréquences de paires et donne un sens à la paire (montre si le mot satellite est le plus souvent cité avec le mot central ou pas)

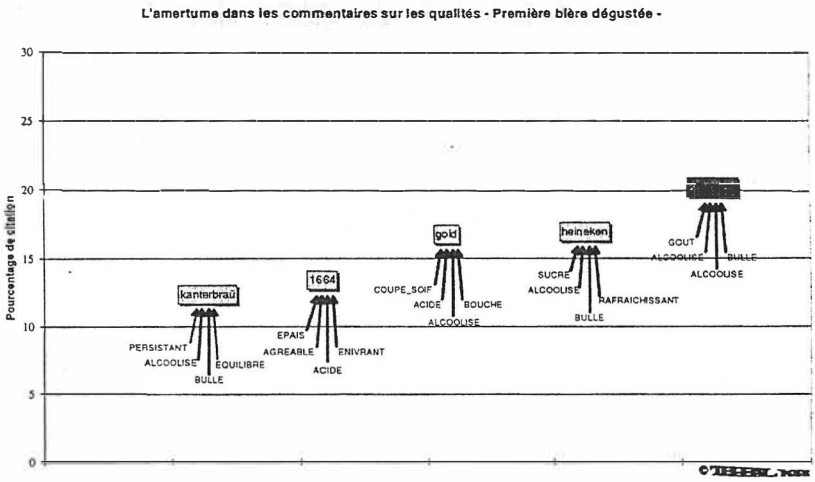
Le choix d'un indice d'association dépend du choix de l'information qu'on a décidé de mettre en évidence.

2.5.3. Choix de la représentation

Nous nous étions orientés au départ sur une représentation en réseau global du type CANDIDE (Teil G., 1994) ou encore MATRISME (BOUTIN E., 1995). Cette approche est très souvent abordée en bibliométrie (ROSTAING H., 1996). De nombreux essais ont montré que ces derniers n'étaient pas adaptés à nos besoins, notamment pour l'interprétation et la comparaison des tests.

Nous avons dû donc mettre au point un nouveau concept que nous avons nommé cartographie de graphes centrés.

2.5.4. Exemple



Les graphes centrés mettent en évidence deux types d'information :

⇒ les termes les plus fortement cités de façon spontanée pour un produit choisi : ils sont en encadré avec une trame claire (pour les moins cités), moyennement claire (pour les moyennement cités) et foncée (pour les plus cités). Ils sont disposés verticalement par ordre de fréquence de citation (ramenée à 100).

Sur l'exemple, le graphe met en évidence les cinq bières les plus citées (Kanterbraü, 1664, Gold, Heineken et 33 Export) pour le mot amertume.

⇒ Les mots satellites (au-dessous des produits) : ils précisent le contexte de citation dans lequel le mot choisi a été cité. Les cinq mots les plus liés au mot choisi apparaissent. L'épaisseur de la flèche indique le degré d'intensité de la liaison. D'autre part, l'orientation de la flèche indique si le mot satellite est toujours associé au mot central ou non.

Sur l'exemple, l'amertume de la Kanterbraü fortement co-cité avec les termes alcoolisé, bulle et moyennement co-citée avec les termes équilibre et persistant.

2. Perspectives, la mise en place d'un questionnaire interactif

Les questions ouvertes nous ont apporté la spontanéité des consommateurs par rapport à leur vision du produit. Nous souhaitons maintenant augmenter le taux de citations des réponses d'une part et aider le consommateur à formuler ses réponses.

Grâce aux réponses issues des questions ouvertes et à l'outil informatique, nous allons mettre en place un questionnaire interactif sur micro-ordinateur. Nous avons choisi de le développer avec des logiciels qui fonctionnent selon les standards d'Internet.

Nous partons du principe que nous connaissons la représentation mentale du produit chez le consommateur grâce à un ensemble de données issues de questions ouvertes. Nous sommes alors en mesure de construire un questionnaire qui préorientera les réponses du consommateur.

Conclusion

Les questions ouvertes nous ont apporté un nouveau type d'information sur les consommateurs. Mais leur exploitation en routine s'avère très difficile (l'étape de postcodage des commentaires est un travail long et complexe qui doit être réalisé pour chaque produit). Nous avons voulu répondre à ces problèmes entre autre en mettant en place le questionnaire interactif d'aide à la formulation des commentaires libres. Il s'agit maintenant de vérifier la concordance des résultats issus d'un test papier avec ceux issus d'un test sur ordinateur.

Références

Chemdata (1996). *TEXTO pour WINDOWS, manuel de référence*, Version 6.0.

Information & Communication (1995). *INFOTRANS Classic Version Française 4.0.* Manuel d'utilisation.

Labbé, D. (1990). Normes de saisie et de dépouillement des textes politiques. *Cahier du C.E.R.A.T.*, n°7, 135p.

Lebart, L, Salem, A. (1994). *Statistique textuelle.* DUNOD, Paris, 342p.

Martin,, N., Rogeaux M. (1994). Etude par analyse textuelle de commentaires de consommateurs après dégustation de boissons. *Sciences des aliments*, 14 (1994), pp. 265-280.

Reinert, M. (1986). Un logiciel d'analyse lexicale : ALCESTE. *Les cahiers de l'analyse des données*, vol. XI, n°4, pp. 471-481.

Rogeaux, M., Ziegelbaum, H. (1996). Comment DANONE prend-il en compte les commentaires sensoriels des consommateurs ? *AGORAL 96*, Lavoisier TEC&DOC, pp. 139-147.

Rostaing, H (1996). *La bibliométrie et ses techniques*. Sciences de la Société, collection « Outils et méthodes », Toulouse, 131p.

Teil, G. (1994). Décrire les goût des fromages : des consommateurs aux experts. *Economie et sociologie rurales*, Grignon 1994, n° 17, vol.1, 2, 3 et 4, Mars 1994.