

---

*Arbres, classes, distances...*

---

*Numéro édité sous la direction de Jacques Poitevineau*  
**Cahiers du LCPE**  
N° 6 - décembre 2002

## **Cahiers du LCPE\***

\*LCPE

44 rue de l'Amiral Mouchez, 75014 Paris

Tél : 01 43 13 56 52

Courrier électronique : [lcpe@ivry.cnrs.fr](mailto:lcpe@ivry.cnrs.fr)

<http://www.ivry.cnrs.fr/lcpe>

Directeur de Publication : Danièle Dubois

Secrétariat de rédaction : Philippe Resche-Rigon

Comité de rédaction :

Sophie David

Anna Maria Lammel

Jacque Poitevineau

## **Sommaire**

<b>Jacques Poitevineau</b> <i>Avant propos</i> .....	5
<b>Patrick Pasquet &amp; Claude Marcel Hladik</b> <i>Utilisation des arbres additifs dans l'étude des covariations des réponses psychophysiques et électrophysiologiques de la perception gustative</i> .....	9
<b>Jacques Poitevineau Bruno Lecoutre, Marie-Paule Lecoutre &amp; Katia Rovira</b> <i>Méthode des arbres de similarité additifs de Sattath et Tversky : illustration dans une tâche de catégorisation de situations d'incertitude</i> .....	15
<b>Sophie David &amp; Catherine Rouby.</b> <i>Groupements d'odeurs ou catégories ? Questions de méthodologie</i> .....	29
<b>Thierry Brenac</b> <i>Méthodes de partition centrale appliquée à l'étude de catégories cognitives</i> .....	71



# Avant Propos

**Jacques Poitevineau**

LCPE/CNRS, ERIS<sup>1</sup>

Les méthodes de représentation de distances, ou de dissemblances, entre des « objets » sous une forme arborée ne sont pas récentes. Elles remontent au moins aux années 60, ne serait-ce qu'avec la publication de l'article de Johnson (1967) dans la revue *Psychometrika*. Dans une telle représentation, les objets correspondent aux nœuds externes (ou feuilles) de l'arbre et la distance entre objets est représentée par la longueur des chemins qui les joignent ou par la hauteur de leur point de jonction. L'arbre peut avoir un intérêt en lui-même, ou ne servir qu'à la construction de classes obtenues en « coupant » certaines branches. En ce qui concerne la méthode de Johnson, il s'agissait de construire une hiérarchie de classes emboîtées les unes dans les autres et formant un arbre ultramétrique. Dans ce cas, si l'on considère deux classes disjointes, les distances intra-classes sont toujours inférieures aux distances inter-classes, et toutes les distances inter-classes sont égales. Ceci est évidemment très contraignant, et a conduit plusieurs auteurs (par exemple, Buneman, 1971) à considérer un modèle moins restrictif, plus apte à représenter les données, et appelé « arbre à distance additive » ou simplement « arbre additif ». En 1977 Sattath et Tversky ont publié dans *Psychometrika* l'algorithme Addtree pour le calcul d'un tel arbre. D'autres algorithmes sont parus depuis (comme NJ, pour Neighbor Joining, de Saitou et Nei, 1987).

Les méthodes d'arbre additif sont couramment utilisées en phylogénie, et aussi en archéologie. On pourrait s'attendre à ce qu'elles le soit également en psychologie cognitive, où, *a priori*, elles paraissent adaptées à rendre compte des phénomènes de catégorisation, et en particulier du concept de prototype (voir, par exemple, Dubois, 1991). Et ce, d'autant que les articles de Johnson et de Sattath et Tversky sont parus dans une importante revue de psychologie. Pourtant, quand on parcourt les années 1996 à 1999 de *Memory & Cognition*, ou les années 1996 et 2000 du *Journal of Experimental Psychology: Learning, Memory & Cognition*, on ne trouve aucun arbre. Ou plutôt, quelques arbres y figurent bien, mais simplement

---

<sup>1</sup>44 rue de l'Amiral Mouchez, 75014 Paris, France, [jacques.poitevineau@ivry.cnrs.fr](mailto:jacques.poitevineau@ivry.cnrs.fr),  
<http://www.ivry.cnrs.fr/lcpe>, ERIS <http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm>

pour représenter la structure *a priori* du matériel expérimental, et en aucun cas comme méthode d'analyse des données. Une possible explication est que ces méthodes sont descriptives (au sens statistique du terme), et ne fournissent donc pas les sacro-saints seuils de signification, encore et toujours (abusivement) considérés comme critères de scientificité, ou tout au moins, comme condition nécessaire à la publication dans une revue scientifique.

Si en phylogénie la pertinence du modèle d'arbre découle de la théorie darwinienne et n'est donc pas remise en cause, dans d'autres domaines se pose la question de la qualité, par rapport aux données recueillies, de la représentation arborée obtenue. Pour y répondre, on dispose d'un certain nombre d'indices d'ajustement (voir, en particulier, Guénoche et Garreta, 2001).

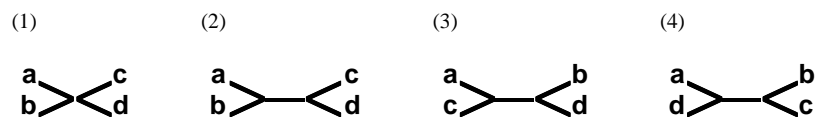
- Ces indices peuvent être métriques ; il s'agit alors de comparer les distances observées et les distances dans l'arbre obtenu. Le *stress* de Kruskal en est un exemple classique, issu des méthodes d'analyse des proximités :

$$\text{stress} = \left( \frac{\sum_i \sum_k [D_A(x_i, x_k) - D(x_i, x_k)]^2}{\sum_i \sum_k D_A(x_i, x_k)^2} \right)^{1/2}$$

où  $D_A(x_i, x_k)$  est la distance, dans l'arbre, entre les objets  $x_i$  et  $x_k$ , et  $D(x_i, x_k)$  la distance (ou dissimilarité) observée.

Plus cette mesure s'approche de zéro, mieux les distances sont reproduites.

- Ils peuvent être topologiques : c'est uniquement la structure de l'arbre qui est alors prise en compte. Parmi les indices de ce type, le taux de quadruplets bien représentés est particulièrement important. Considérons un quadruplet quelconque, par exemple les 4 objets a, b, c, d. Les sommes des distances deux à deux informent sur la topologie sous-jacente qui est forcément l'une des suivantes :



Par exemple, si  $D(a,b)+D(c,d)$  est la plus petite des sommes, c'est la topologie (2) qui est en jeu. Le taux de quadruplets bien représentés est la proportion de quadruplets qui ont la même topologie selon la distance "observée" (celle provenant des données) et la distance d'arbre. Bien évidemment, on sera d'autant plus satisfait que ce taux s'approche de 100%.

Le but de ce numéro est de présenter quelques exemples d'utilisation des arbres additifs, avec essentiellement une visée méthodologique. Ces exemples proviennent de champs variés :

- Il s'agit d'abord de l'électrophysiologie et de la psychophysique, avec l'étude de Pasquet et Hladik sur l'évolution de la perception gustative, avec une comparaison de données obtenues sur des primates non humains et de données sur des sujets humains.

- Dans l'article suivant, exploration du concept de hasard, des sujets de différents degrés d'expertise en probabilité doivent catégoriser diverses situations d'incertitude. A cette occasion, on verra comment il est possible de modifier la procédure en introduisant des objets fictifs dans le but de différencier les sujets selon le sens que ceux-ci attribuent aux classes qu'ils constituent.

- Dans l'article de David *et al.*, traitant de la qualité hédonique de certains odorants, plusieurs expériences sont décrites. L'accent est mis sur l'utilisation des critères d'ajustement, à la fois dans l'étude d'une expérience particulière, et dans la comparaison d'une expérience à une autre. On verra aussi comment les fondements de l'organisation catégorielle étudiée sont objectivés par le croisement des topologies obtenues avec les verbalisations des sujets.

Ces trois articles ont en commun de comparer des arbres entre eux. Cette comparaison peut se faire qualitativement, mais il est aussi possible de la quantifier en utilisant, par exemple, la distance topologique de Robinson et Foulds. Dans l'article sur le Hasard, celle-ci est utilisée entre les arbres « observés », c'est-à-dire issus de l'expérimentation, mais aussi entre ceux-ci et un arbre « théorique ».

Enfin, le dernier article s'inscrit dans une perspective différente puisqu'il offre une autre approche de traitement que celle des arbres. En effet, comme c'est le cas dans les deuxième et troisième articles, les données de base sont souvent constituées par des partitions des objets, réalisées par les sujets, par exemple au cours d'une épreuve de tri libre (les sujets sont invités à regrouper les objets en un nombre de « tas » à leur convenance). Et il se peut que la question centrale du chercheur ait trait aux classes d'objets, indépendamment des relations entre elles. Alors, au lieu de construire un arbre pour en déduire ensuite les classes, on peut tout naturellement chercher plus directement une partition « moyenne » de celles réalisées par les sujets. C'est à ce problème que répondent les méthodes de « partition centrale », et que Thierry Brenac nous fait découvrir. Il illustre son propos par une étude de la catégorisation de voies routières urbaines.

On espère que ce numéro contribuera à montrer que les données sous forme de distances entre des « objets » ne relèvent pas seulement de l'analyse des proximités (*MultiDimensional Scaling* en anglais) dans lesquelles les objets sont positionnés dans un espace à plusieurs dimensions muni d'une certaine métrique (euclidienne le plus couramment), et que les analyses arborées sont non seulement faisables, mais aussi intéressantes.

## Programmes

Pour pouvoir réellement utiliser ces méthodes, il faut bien sûr disposer d'un programme informatique.

Au LCPE, la version d'Addtree de Barthélémy et Guénoche (1991), version remaniée de l'algorithme original, est la plus utilisée. Elle présente le grand avantage de fournir un grand nombre de critères de la qualité d'ajustement du modèle d'arbre aux données, à la fois métriques et topologiques.

Une autre version d'Addtree, ainsi que d'autres programmes, sont disponibles dans le logiciel T-Rex, que l'on peut se procurer gratuitement à l'adresse

suivante : <http://www.fas.umontreal.ca/BIOL/Casgrain/fr/labo/>. Les programmes d'Alain Guénoche figurent aussi à cette adresse, de même que le calcul de la distance de Robinson et Foulds entre arbres.

Le logiciel statistique américain Systat (version 9) offre également une version d'Addtree.

Quant aux méthodes de partition centrale, il est plus difficile d'y avoir accès. Thierry Brenac a écrit des programmes, mais leur exécution nécessite de posséder le logiciel Matlab 6.

## ***Références***

- Barthélémy J.P., Guénoche A. (1991) *Trees and Proximity Representations*, New York, Wiley.
- Buneman P. (1971) The recovery of trees from measures of dissimilarity, in F.R. Hodson, D.G. Kendall, P. Tautu (éds.) *Mathematics in the Archaeological and Historical Sciences*, Edinburgh, Edinburgh University Press, 387-395.
- Dubois D. (éd.) (1991) *Sémantique et cognition. Catégories, prototypes, typicalité*, Paris, éditions du CNRS.
- Guénoche A., Garreta H. (2001) Can we have confidence in a tree representation ? *Proceedings of JOBIM'2000, Lecture Notes in Computer Sciences*, vol. 2066, 43-53.
- Johnson S.C. (1967) Hierarchical clustering schemes, *Psychometrika*, 32, 241-254.
- Sattath S., Tversky A. (1977) Additive similarity trees, *Psychometrika*, 42, 319-345.
- Saitou N., Nei M. (1987) The neighbor-joining method, a new method for reconstructing phylogenetic trees, *Molecular Biology Evolution*, 4, 406-425.



# Utilisation des arbres additifs dans l'étude des covariations des réponses psychophysiques et électrophysiologiques de la perception gustative

***Patrick Pasquet, Claude Marcel Hladik***

*Dynamique de l'Evolution Humaine, CNRS UPR 2147, F75014  
Eco-Anthropologie, CNRS FRE2323, Muséum National d'histoire Naturelle, F75005*

Nous avons utilisé une même méthode d'analyse des corrélations par les arbres additifs (Sattath et Tversky, 1977), afin de comparer, dans une perspective d'étude de l'évolution de la perception gustative, les covariations des réponses électrophysiologiques obtenues avec des primates non humains et celles obtenues par l'approche psychophysique sur des échantillons de populations humaines .

Les études sur les primates non humains concernent le ouistiti (*Callithrix jacchus*), le macaque rhesus (*Macaca mulatta*) et le chimpanzé (*Pan troglodytes*). Les enregistrements des impulsions sur 50 fibres isolées du nerf gustatif (fréquence des impulsions sur les fibres isolées du nerf de la gustation) ont été réalisées à l'Université de Wisconsin, après stimulation par différents produits de la langue des animaux anesthésiés (Hellekand é Ninomya, 1994 ; Hellekant *et al.*, 1997 a, 1997b, 1998 ; Danilova *et al.*, 1998).

Chez l'humain, les données concernent les seuils de reconnaissance de différentes solutions de produits purs (fructose, saccharose, chlorure de sodium, acide citrique, acide tannique, tannin de chêne, quinine et 6-n-propylthiouracil) qui représentent une sélection des produits testés sur les primates non humains. Les tests ont été réalisés en simple aveugle, la reconnaissance verbale permettant de définir le seuil au cours d'une séance de présentation des solutions à concentration croissante (Hladik *et al.*, 1986, 2002). L'ensemble des sujets testés forme un échantillon composite de 412 adultes de 18 à 59 ans, issu de différentes populations d'Europe et d'Afrique.

Le programme SYSTAT 9 (SPSS Inc., Chicago) a été utilisé pour le calcul des matrices de distances à partir des corrélations entre les réponses aux différents produits testés. Le programme T-Rex disponible gratuitement sur le site [www.fas.umontreal.ca/BIOL/Casgrain/en/labo/t-rex/index.html](http://www.fas.umontreal.ca/BIOL/Casgrain/en/labo/t-rex/index.html) a permis d'obtenir une représentation radiale des arbres additifs issus des matrices de distances.

Les résultats obtenus avec des produits solubles analogues (Fig. 1 et 2) montrent que les similarités et dissimilarités entre les réponses pour ces différents produits

sont très comparables d'une espèce à l'autre, en dépit de la spécificité des méthodes utilisées pour l'approche psychophysique et l'électrophysiologie.

Chez l'homme (Fig. 1), la corrélation entre la perception des deux sucres (fructose et saccharose) indique que les signaux perçus ne diffèrent que très légèrement ; cependant le niveau modéré de la corrélation ne permet pas d'affirmer l'existence d'un type unique de récepteur du « goût sucré ». Dans le cas des perceptions du goût amer de la quinine et du goût des tannins — un goût astringent qui est plus qu'un simple phénomène tactile (Hellekant *et al.*, 1993) —, les corrélations significatives mettent en évidence que plusieurs récepteurs périphériques sont communs à chacune de ces perceptions. L'absence de covariation entre l'ensemble des tannins+quinine et celui des sucres montre que, parmi les transducteurs gustatifs mis en jeu pour chacun de ces groupes de substances, il n'y a pas de récepteur commun.

L'introduction, dans l'analyse, des données sur le cas particulier de la perception du 6-*n*-propylthiouracile (PROP), perçu comme amer par certains « goûteurs » (Bartoshuk, 1979) n'affecte pas la structure générale de l'arbre additif, ajoutant une branche indépendante (non représentée sur la fig. 1), qui témoigne de l'absence de covariation avec la perception des autres substances incluant le groupe quinine+tannins.

Chez les primates non humains (Fig. 2), les corrélations entre les signaux enregistrés sur les fibres du nerf de la gustation montrent également une dichotomie entre les perceptions : d'une part les fibres du nerf gustatif qui réagissent aux sucres présentent des différences entre fructose et saccharose (particulièrement nettes chez le macaque) ; d'autre part, les réactions aux tannins et aux alcaloïdes covarient chez toutes les espèces testées (ouistiti, macaque et chimpanzé). Dans tous les cas, la corrélation négative entre ces deux ensembles de signaux, correspondant respectivement à des substances bénéfiques (les sucres, source d'énergie) et à des produits que les primates doivent éviter (à effet toxique, comme beaucoup d'alcaloïdes, ou anti-nutriments, comme les tannins), montre qu'il n'y a pas de récepteurs périphériques communs à ces deux ensembles de réponses gustatives.

Cette dichotomie apparaît comme le résultat de la co-évolution des primates et des angiospermes dont les fruits sucrés permettent la dissémination des graines, mais dont les teneurs en produits secondaires (alcaloïdes, tannins, terpènes, etc.) nécessitent une détection rapide et leur évitement grâce au réflexe gusto-facial (Steiner *et al.*, 2001). L'appareil gustatif des primates permet de détecter de nombreuses autres substances, en particulier le chlorure de sodium. Cependant il n'apparaît pas que le goût salé, tel qu'il est perçu et décrit par l'homme, résulte de la co-évolution avec des aliments potentiels des milieux naturels dont les teneurs en chlorure de sodium sont inférieures au seuil de détection (Hladik et Simmen, 1996). L'absence de covariation des signaux déclenchés par le sel avec les deux grands ensembles de perceptions (sucres et tannins+alcaloïdes), ainsi que les teneurs en chlorure de sodium des aliments naturels inférieures aux seuils de gustation des primates, remettent en cause la notion de « goûts de base ». Bien

que l'importance du sel dans les civilisations montre que son intégration dans une alimentation élaborée est une adaptation culturelle qui aurait pu se manifester dès l'émergence des premiers hominidés, la détection des solutions concentrées de chlorure de sodium (le goût salé) ne serait — comme dans le cas du PROP — que le résultat fortuit de la co-évolution d'un ensemble de gènes permettant aux primates une réaction rapide aux produits toxiques et aux substances énergétiques.

## Références

- Bartoshuk L. M. (1979) Bitter taste of saccharin related to the genetic ability to taste the bitter substance 6-n-propylthiouracyl, *Science*, 206, 934-935.
- Danilova V., Hellekant G., Roberts T., Tinti J.-M., Nofre C. (1998) Behavioral and single chorda tympani taste fiber responses in the common marmoset, *Callithrix jacchus jacchus*, *Annals N. Y. Acad. Sci.*, 855, 160-164.
- Hellekant G., Danilova V., Ninomiya, Y. (1997a) Primate sense of taste: behavioral and single chorda tympani and glossopharyngeal nerve fibers recordings in the rhesus monkey, *Journal of Neurophysiology*, 77, 978-993.
- Hellekant G., Ninomiya Y. (1994) Bitter taste in single chorda tympani taste fibers on chimpanzee, *Physiol. Behav.*, 56, 1185-1188.
- Hellekant G., Ninomiya Y., Danilova V. (1997b) Taste in chimpanzees II, single chorda tympani fibers, *Physiol. Behav.*, 61, 829-841.
- Hellekant G., Ninomiya Y., Danilova, V. (1998) Taste in chimpanzees. III, Labeled-line coding in sweet taste, *Physiol. Behav.*, 65, 191-200.
- Hellekant G., Hladik C.M., Dennys V., Simmen B., Roberts T.W., DuBois G., Walter D. E. (1993) On the sense of taste in two Malagasy primates (*Microcebus murinus* and *Eulemur mongoz*), *Chem. Senses*, 18, 307-320.
- Hladik C. M., Pasquet P. Simmen B. (2002) New perspective on taste and primate evolution: the dichotomy in gustatory coding for perception of beneficent vs. noxious substances as supported by correlations among human thresholds, *Am. J. Phys. Anthropol.*, 117, 342-348.
- Hladik C. M., Robbe B., Pagezy H. (1986) Sensibilité différentielle des populations Pygmées et non Pygmées de forêt dense, de Soudaniens et d'Eskimos, en rapport avec l'environnement biochimique, *C. R. Acad. Sci., Paris*, 303, 453-458.
- Hladik C. M., Simmen B. (1996) Taste perception and feeding behavior in nonhuman primates and human populations, *Evolutionary Anthropology*, 5, 58-71.
- Sattath S., Tversky A. (1977) Additive similarity trees, *Psychometrika*, 42, 319-345.
- Steiner J. E., Glaser D., Hawilo M. E., Berridge K.C. (2001) Comparative expression of hedonic impact; affective reactions to taste by human infants and other primates, *Neuroscience and Behavioral Reviews*, 25, 53-74.

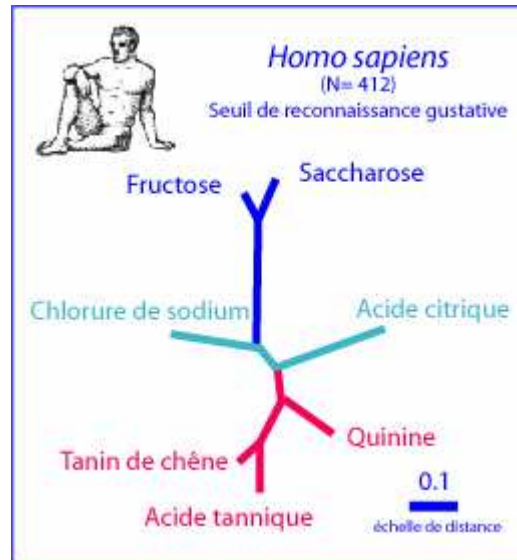


Figure 1. Arbre additif (stress = 2,5 %) montrant la relation entre les seuils de reconnaissance gustatifs pour différents produits, chez 412 sujets humains.

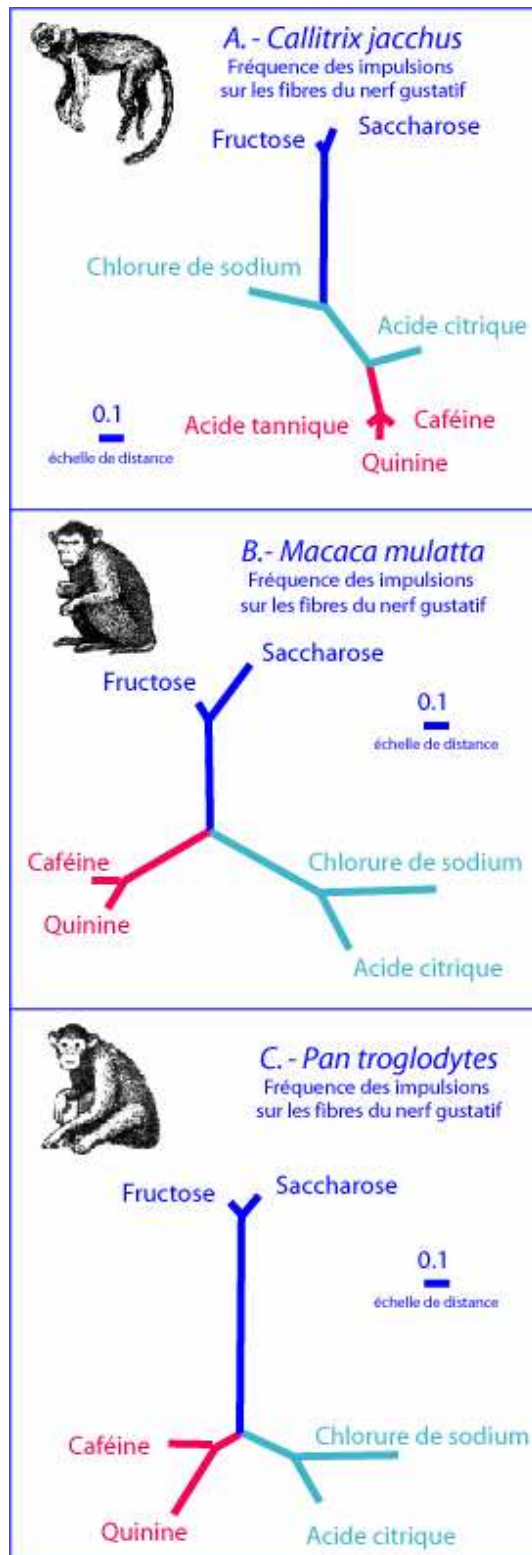


Figure 2. Arbres additifs montrant les relations entre les signaux enregistrés sur les fibres isolées du nerf gustatif de trois espèces de primates non humains. A : ouistiti (*Callithrix jacchus*), stress = 1,5 %. B : macaque rhésus (*Macaca mulatta*), stress = 5,0 %. C : chimpanzé (*Pan troglodytes*), stress = 3,5 %.



# Méthode des arbres de similarité additifs de Sattath et Tversky : Illustration dans une tâche de catégorisation de situations d'incertitude<sup>1</sup>

*Jacques Poitevineau*<sup>(a,b)</sup>, *Bruno Lecoutre*<sup>(a,c)</sup>,  
*Marie Paule Lecoutre*<sup>(a,d)</sup>, *Katia Rovira*<sup>(d)</sup>

<sup>a</sup>ERIS

<sup>b</sup>LCPE, C.N.R.S

<sup>c</sup>Laboratoire de Mathématiques Raphaël Salem<sup>2</sup>

<sup>d</sup>Laboratoire Psy.co, E.A. 1780, Université de Rouen<sup>3</sup>

## Introduction

Dans une tâche de catégorisation d'objets, les données à analyser se présentent sous forme d'un ensemble de partitions des objets. A partir de cet ensemble on peut facilement construire une matrice de dissimilarités entre les objets. La méthode des arbres de similarité additifs de Sattath et Tversky (1977) permet de représenter la structure des objets sous la forme d'un arbre (un ensemble de nœuds reliés par des arêtes). Les objets correspondent alors aux nœuds externes (ou feuilles) de l'arbre, et la dissimilarité entre deux objets est représentée par la longueur du chemin qui les joint. Le programme informatique Addtree (dans la version de Barthélemy et Guénoche, 1991) permet de construire ces arbres.

Nous illustrerons ici l'utilisation de cette méthode pour analyser les données obtenues dans une tâche de catégorisation de situations d'incertitude par des sujets appartenant à trois groupes se différenciant selon leur degré d'expertise en

---

<sup>1</sup> Marie-Paule Lecoutre et Katia Rovira sont les auteurs de l'expérience présentée ici pour illustrer la méthode ; Jacques Poitevineau et Bruno Lecoutre sont plus particulièrement responsables des analyses statistiques.

<sup>2</sup> UMR 6085, C.N.R.S. et Université de Rouen Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex,  
bruno.lecoutre@univ-rouen.fr, <http://www.univrouen.fr/LMRS>

<sup>3</sup> UFR Psychologie, Sociologie, Sciences de l'Éducation 76821 Mont-Saint-Aignan Cedex,  
marie-paule.lecoutre@univ-rouen.fr, <http://www.univ-rouen.fr/psy-socio-sceduc/People>  
katia.rovira@univ-rouen.fr

probabilités. Les représentations par des arbres sont utilisées ici pour faire ressortir la structure des objets et la comparer à une classification théorique *a priori*. La méthode est également appliquée pour rechercher une classification des sujets à l'intérieur de chacun des groupes.

## **I. Une expérience de catégorisation de situations d'incertitude**

### **I.1. Problématique de l'expérience**

Le hasard est une notion particulièrement complexe et ambiguë qui a donné et continue de donner lieu à de nombreuses interprétations. Au 19<sup>ème</sup> siècle, en accord avec la conception Laplacienne déterministe du monde, le hasard est le nom donné à l'ignorance d'une personne dans un univers entièrement déterminé. Au contraire, quelques théories plus récentes reposent sur l'idée que le hasard est dans l'essence même des choses. Par exemple, en physique quantique le comportement des particules est conçu comme fondamentalement probabiliste. Pour formaliser la notion de hasard le concept de probabilité a été utilisé dès le 17<sup>e</sup> siècle par les mathématiciens. De nos jours la communauté scientifique s'accorde à reconnaître deux interprétations principales. (1) La plus classique est une interprétation dite *fréquentiste* ou *objectiviste* selon laquelle la probabilité renvoie à des fréquences objectives de réalisation d'événements. (2) Il existe une interprétation plus large dite *subjectiviste* selon laquelle la probabilité traduit un degré de confiance dans la vérité d'une proposition ; c'est à cette dernière interprétation que se réfère la théorie statistique Bayésienne. L'étude expérimentale présentée ici a pour objectif d'apporter des éléments de réponse aux questions suivantes. Que mettent spontanément les gens derrière le mot « hasard » ? Existe-t-il des interprétations dominantes ? Dans quelle mesure sont-elles liées au degré d'expertise en probabilité ? Sont-elles proches de certaines conceptualisations probabilistes ?

### **I.2. Méthode**

#### ***Sujets***

Pour étudier l'effet du degré d'expertise en probabilités nous avons observé trois groupes de sujets: (1) 20 enseignants-chercheurs en mathématiques considérés comme experts ; (2) 20 enseignants-chercheurs en psychologie confrontés à la notion de hasard dans leur pratique statistique ; (3) 20 collégiens en classe de 3<sup>ème</sup> n'ayant pas reçu d'enseignement en probabilités.

#### ***Matériel***

La situation expérimentale consiste en une tâche de catégorisation d'évènements incertains. Nous considérons que la façon dont un sujet classe un matériel permet



d'avoir accès à ses représentations (cf. par exemple Medin, Lynch et Coley, 1997). Le matériel comprend 16 situations d'incertitude (présentées dans le tableau I) écrites sur des cartons blancs que le sujet peut manipuler. Le choix des situations s'appuie notamment sur les travaux de Nisbett *et al.* (1983) et Konold *et al.* (1991). Il y a 8 situations « réelles » et 8 situations « stochastiques ». Les situations réelles sont inspirées d'événements de la vie courante (par exemple le fait d'attraper un rhume dans le courant du mois prochain). Dans 4 situations il y a implication du sujet, ce dernier étant désigné par « vous » dans l'énoncé ; dans les 4 autres il n'y a pas d'implication, le sujet n'étant pas désigné dans l'énoncé. Les situations stochastiques renvoient à des processus répétables (par exemple le lancer d'un dé) ou des situations de tirages au sort. Quatre situations ont des issues symétriques c'est-à-dire équiprobables ; les 4 autres ont des issues non symétriques.

**situations « réelles »**

*Avec implication explicite du sujet*

- A « Ami » Le fait de rencontrer un ami que vous n'avez pas vu depuis 10 ans
- B « Loto » Le fait que vous gagniez 10000F au loto
- C « Tête » Le fait de dire la première chose qui vous passe par la tête
- D « Rhume » Le fait que vous attrapiez un rhume dans le courant du mois prochain

*Sans implication explicite du sujet*

- E « Graine » Le fait qu'une graine mise en terre germe
- F « Bourse » Le fait que le cours d'une action à la bourse de Paris aura progressé de plus de 5% dans trois mois
- G « Pluie 1 » Le fait qu'il a plu à Paris le 15 mars 1936
- H « Pluie 2 » Le fait qu'il pleuve demain à Paris

**situations « stochastiques »**

*Avec issues symétriques*

- S « Dé » Le fait d'obtenir un chiffre pair à l'issue du lancer d'un dé
- T « Pièce » Le fait d'obtenir Pile à l'issue du lancer d'une pièce de monnaie non truquée
- U « Faces » Le fait d'obtenir Face au 5<sup>ème</sup> lancer d'une pièce de monnaie non truquée qui est tombée sur Face les quatre fois précédentes
- V « Boules 1 » Le fait de tirer une boule blanche d'une boîte qui contient 10 boules noires et 10 boules blanches

*Avec issues non symétriques*

- W « Jetons » Le fait de tirer simultanément 2 jetons rouges d'une boîte qui contient 1 jeton blanc et 2 jetons rouges
- X « Chaussettes » Le fait de constituer une paire de chaussettes « assorties » à partir d'un tirage à l'aveugle de deux chaussettes d'un tiroir qui contient 2 paires de chaussettes différentes
- Y « Bonbons » Le fait de tirer un bonbon au citron d'une boîte qui contient 20 bonbons à l'orange et 10 au citron
- Z « Boules 2 » Le fait de tirer une boule blanche d'une boîte qui contient 10 boules noires et 20 boules blanches

Tableau I - Les 16 situations d'incertitude considérées

### ***Procédure***

L'expérience se déroule en deux phases. (1) On demande au sujet de mettre ensemble les situations qui se ressemblent, il peut faire autant de classes qu'il le souhaite ; il s'agit d'une tâche de « catégorisation libre » dans laquelle la notion de hasard n'est pas explicitée. (2) On reprend toutes les situations et le sujet doit dire, pour chacune d'elle, si le hasard intervient ou non ; il s'agit d'une tâche de « catégorisation contrainte » dans laquelle la notion de hasard est alors explicite. Dans les deux cas il est demandé au sujet de justifier ses classements et ses réponses.

## **II. Résultats**

Les résultats ont été analysés avec plusieurs méthodes complémentaires et les interprétations données dans cet article s'appuient sur l'ensemble de ces méthodes. Nous ne présenterons ici que la méthode des arbres de similarité additifs, qui, comme nous l'avons dit dans l'introduction, permet à partir des catégorisations observées de construire : (1) des arbres associés aux situations pour étudier leur structure ; (2) des arbres associés aux sujets pour rechercher une classification de ces derniers.

### **II.1. Ajustement par un modèle d'arbre**

Pour juger de la qualité d'une représentation par un arbre, on dispose d'un ensemble d'indices (*cf.* Guénoche et Garreta, 2001 et l'avant propos de ce numéro pour des définitions), qui se regroupent en deux catégories. (1) Les indices métriques comparent les distances déduites des partitions effectuées par les sujets avec les distances dans la représentation arborée. (2) Les indices topologiques ne font intervenir que la structure de l'arbre.

Il est également intéressant de considérer le rapport des longueurs des arêtes internes à celles des arêtes externes (un arbre est plus facile à reconstituer quand ce rapport est grand). Ce rapport ne fournit pas un critère d'ajustement, mais il joue un rôle important dans les études de simulation qui servent à avoir une idée de la variation des indices quand les distances sont « bruitées » aléatoirement.

Le tableau II donne, pour les arbres associés aux situations, la valeur de ce rapport et de deux indices de qualité de représentation : un indice métrique, le stress de Kruskal, et un indice topologique, le taux de quadruplets bien représentés. La représentation est d'autant meilleure que le stress est plus petit et que le taux de quadruplets est élevé.

	Rapport arêtes internes/externes	Stress	Taux de quadruplets bien représentés
<b>Catégorisation libre</b>			
Mathématiciens	0,45	0,06	87%
Psychologues	0,59	0,08	87%
Collégiens	0,43	0,13	85%
<b>Catégorisation contrainte</b>			
Mathématiciens	0,80	0,08	90%
Psychologues	0,56	0,12	82%
Collégiens	0,73	0,12	71%

Tableau II – Indices de la qualité de représentation des arbres associés aux situations

En se référant aux études de simulation effectuées par Guénoche et Garreta (2001), les valeurs obtenues permettent de considérer qu'un modèle d'arbre apparaît raisonnablement ajusté aux données dans tous les cas. La meilleure représentation est obtenue pour les mathématiciens qui ont le plus petit stress et le plus grand taux de quadruplets bien représentés. En ce qui concerne les arbres associés aux sujets, les indices (non présentés ici) sont moins bons mais restent encore acceptables. Nous commentons ci-après les arbres obtenus dans chacune des deux phases.

## II.2. Première phase : Catégorisation libre (hasard non explicite)

### *Arbres associés aux situations*

Les trois structures d'arbres pour les trois groupes sont voisines et reflètent assez bien la classification théorique à savoir : la dissociation entre situations réelles et stochastiques, et à l'intérieur de celles-ci la différence entre implication/sans implication du sujet et symétrie/non-symétrie des issues. Pour les mathématiciens, comme le montre la figure 1, on retrouve même exactement cette classification.

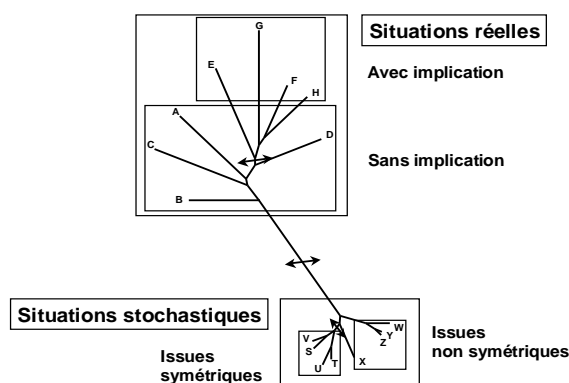


Figure 1 – Mathématiciens : arbre associé aux situations dans la catégorisation libre

Pour les collégiens (voir la figure 2), nous retrouvons les deux classes principales mais la situation B (Loto) est classée avec les situations stochastiques. Pour les psychologues, la structure de l'arbre (non reproduit ici) est intermédiaire.

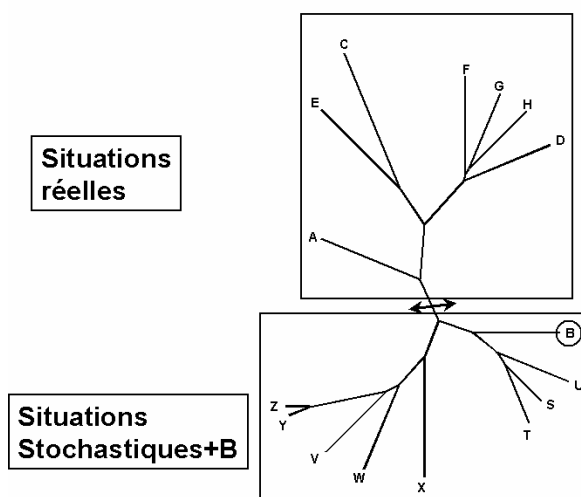


Figure 2 – Collégiens : arbre associé aux situations dans la catégorisation libre

*Comparaison des arbres associés aux situations*

Pour comparer les arbres, on a calculé la distance de Robinson et Foulds (1981) entre ces arbres (cf. le tableau III). Cette distance est purement topologique, c'est-à-dire qu'elle ne tient compte que de la structure des arbres et pas de la longueur des arêtes. Elle est égale au nombre minimum d'opérations élémentaires (fusion ou division de nœuds) nécessaires pour transformer un arbre en un autre et est comprise entre 0 et 26 (c'est-à-dire  $2k-6$ ,  $k=16$  étant le nombre d'objets). Par ailleurs, on a également calculé les distances avec l'arbre théorique correspondant à la structure *a priori* des situations (figure 3).

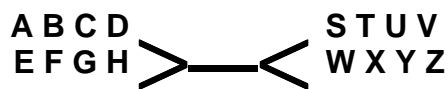


Figure 3 – Arbre théorique associé aux situations

	Théorique	Mathématiciens	Psychologues
Mathématiciens	12		
Psychologues	16	22	
Collégiens	18	20	12

Tableau III - Distance de Robinson et Foulds entre les arbres des situations

On vérifie que l'arbre des mathématiciens est le plus proche de l'arbre théorique, et que les arbres des collégiens et des psychologues, relativement proches entre eux, en sont à peu près équidistants. On pourrait s'étonner du fait que la distance entre l'arbre théorique et l'arbre des mathématiciens n'est pas nulle, alors qu'on a indiqué que les classifications des situations étaient identiques dans les deux cas. Ceci provient de la « structure fine » de l'arbre des mathématiciens : la distance de 12 est entièrement due aux petites arêtes qui distinguent les situations appartenant à une même catégorie *a priori*.

### Arbres associés aux sujets

Les arbres associés aux sujets permettent d'affiner les résultats globaux précédents. Chez les mathématiciens (figure 4), seuls 3 sujets réalisent la même classification. On peut cependant identifier 3 sous-groupes de sujets: (1) 7 sujets qui regroupent la majorité des situations dans une seule classe ; (2) 7 sujets qui séparent les situations réelles des situations stochastiques ; (3) 6 sujets qui répartissent les situations dans un nombre élevé de classes (entre 4 et 12).

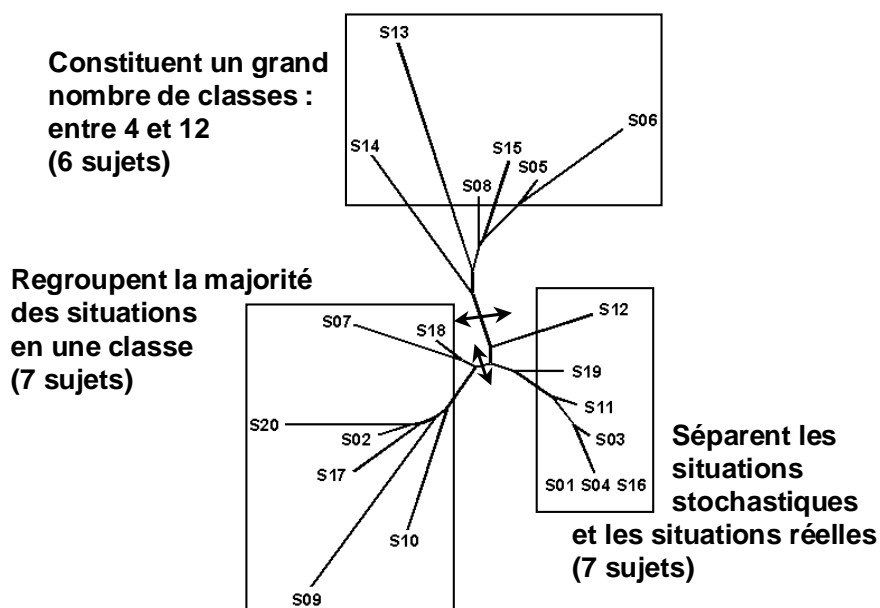


Figure 4 – Mathématiciens : arbre associé aux sujets dans la catégorisation libre

Chez les Psychologues (figure 5) on observe une structure d'arbre nettement différente. (1) 6 sujets regroupent en une classe les 8 situations stochastiques ; (2) 3 sujets regroupent dans une classe les 8 situations stochastiques avec en plus des situations réelles ; (3) 8 sujets constituent un grand nombre de classes (entre 4 et 7). Seuls deux sujets réalisent la même classification, à savoir une classe unique ; un sujet réalise deux classes.

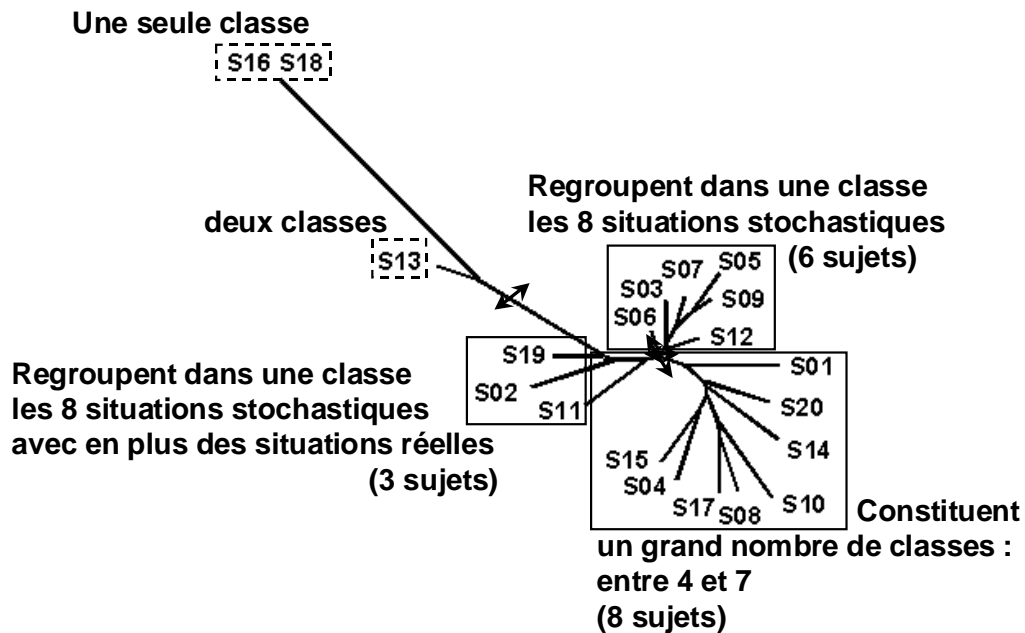


Figure 5 – Psychologues : arbre associé aux sujets dans la catégorisation libre

Chez les collégiens (figure 6) nous notons encore une structure d’arbre différente et autant de classifications qu’il y a de sujets. Il y a également 3 sous-groupes de sujets mais avec des interprétations différentes : (1) 8 sujets regroupent dans une classe au moins 6 situations stochastiques ; (2) 7 sujets regroupent dans une classe 4 situations réelles ; (3) 5 sujets mélangeant les deux types de situations.

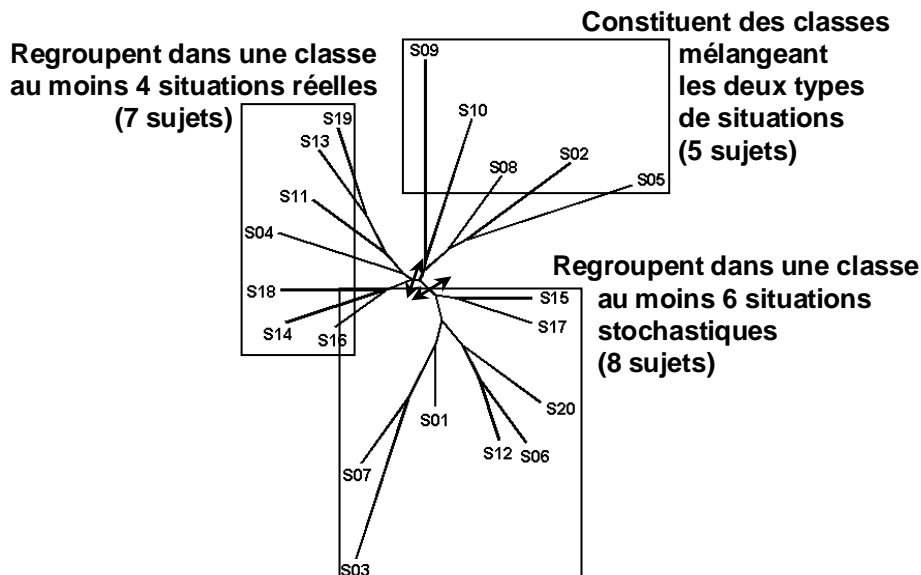


Figure 6 – Collégiens : arbre associé aux sujets dans la catégorisation libre

Si l'on considère les trois arbres, on peut noter une très grande variabilité interindividuelle dans chacun des trois groupes et une variabilité d'un groupe à l'autre, puisque la structure des arbres diffère selon les groupes. Ces résultats montrent que les sujets se différencient dans leur façon de catégoriser les situations et donc selon les critères qu'ils utilisent.

### **II.3. Deuxième phase : Catégorisation contrainte (hasard explicite)**

Dans la situation de catégorisation contrainte, le sujet doit dire si pour chacune des 16 situations le hasard intervient ou non. Les réponses ont conduit à considérer deux classes supplémentaires, l'une *intermédiaire* qui regroupe des réponses nuancées du type « le hasard intervient un peu moins », et l'autre qui correspond à quelques cas où le sujet déclarait qu'il lui était impossible de se prononcer. Ainsi le nombre de classes des catégorisations analysées varie de 1 à 4. Cette situation pose en fait un problème particulier, puisqu'on sait ici qu'une même catégorisation observée correspond dans certains cas à deux réponses opposées. Ainsi par exemple, quelques sujets répondent que le hasard intervient dans toutes les situations, alors que d'autres répondent qu'il n'intervient dans aucune. Dans ces deux cas cela revient à regrouper toutes les situations en une seule classe. Plus généralement deux sujets qui ont donné des réponses diamétralement opposées (pour chaque situation, si un sujet l'a qualifiée de « hasard », l'autre l'a qualifiée de « pas de hasard ») sont considérés comme « équivalents ». Si on applique la méthode directement aux partitions correspondant aux réponses des sujets, on ne tient pas compte de la nature de la réponse.

Nous avons néanmoins directement appliqué la méthode dans un premier temps. Nous avons ensuite envisagé la modification suivante. Quatre objets *fictifs*, « hasard », « pas de hasard », « intermédiaire » et « ? » ont été ajoutés. Dans chacune des classes d'objets constituées par un sujet, on ajoute l'objet fictif correspondant à la qualification de la classe. En outre, si le sujet n'a pas lui-même créé toutes les 4 classes, on ajoute les classes manquantes qui ne contiennent alors que l'objet fictif correspondant. Ainsi, deux sujets réalisant la même partition mais avec des qualifications différentes sont distingués. Il est important de noter que cet ajout d'objets fictifs ne change pas les distances entre les « vrais » objets (non fictifs). Par construction, les 4 objets fictifs seront toujours disjoints et la distance entre eux sera maximale. En outre, ces objets facilitent l'interprétation des classes dans les arbres associés aux situations.

#### ***Arbres associés aux situations***

##### *Application directe de la méthode*

Les trois structures d'arbres associés aux situations sont encore voisines pour les trois groupes de sujets, et on retrouve bien la distinction entre situations stochastiques et réelles. Cependant les deux situations réelles A (Ami) et B (Loto) apparaissent intermédiaires pour les mathématiciens (comme on peut le voir sur la figure 7), et même nettement plus proches des situations stochastiques que des

réelles pour les deux autres groupes (ce qui pour les collégiens est conforme au résultat obtenu dans la catégorisation libre). Par ailleurs, on ne retrouve plus, même chez les mathématiciens, la distinction entre les sous-catégories définies a priori par l'implication du sujet et par le type d'issues.

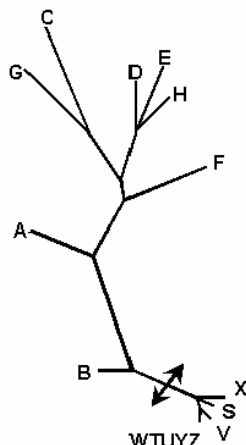


Figure 7 – Mathématiciens : arbre associé aux situations dans la catégorisation contrainte (application directe de la méthode)

#### Ajout d'objets fictifs

La classification des situations est peu modifiée (voir la figure 8 pour les mathématiciens) de même que les indices de qualité de représentation. Dans les 3 groupes les situations stochastiques sont proches du hasard et les situations réelles plus proches de l'absence de hasard, à l'exception de A et B. Chez les mathématiciens, c'est la situation C (dire ce qui vous passe par la tête) qui est la plus proche de l'absence de hasard ; chez les psychologues et les collégiens, c'est d'abord la situation E (Graine) puis la situation C.

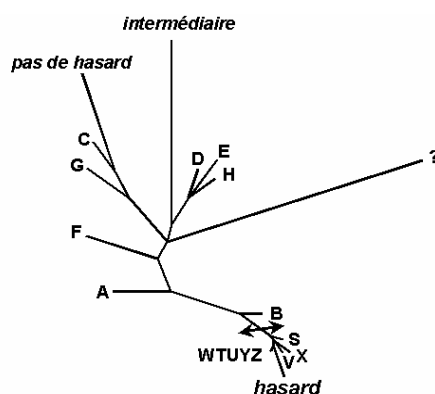


Figure 8 – Mathématiciens : arbre associé aux situations dans la catégorisation contrainte (avec ajout d'objets fictifs)



### Arbres associés aux sujets

Les arbres associés aux sujets reflètent une variabilité inter-individuelle (et inter-groupe) encore plus importante pour la catégorisation avec contrainte que pour la catégorisation libre. On peut le voir pour les mathématiciens sur la figure 9 pour l'application de la méthode avec ajout d'objets fictifs.

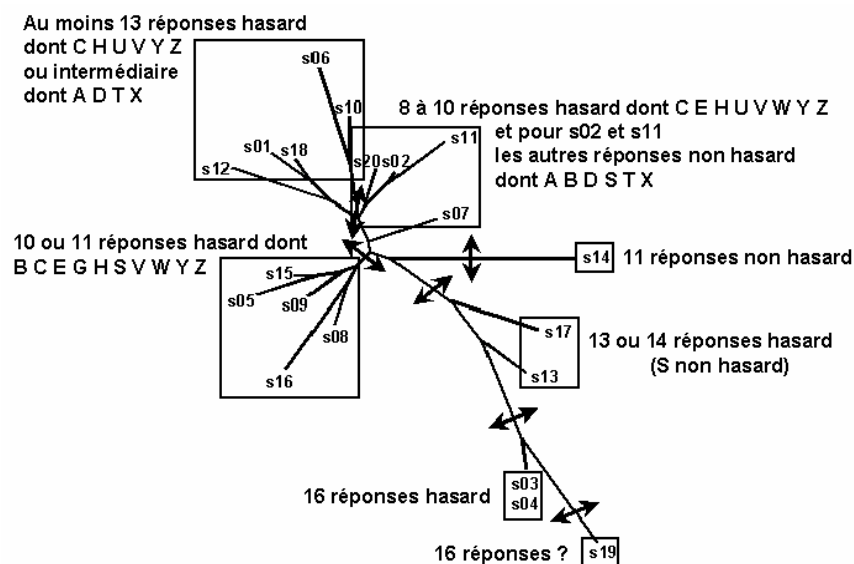


Figure 9 – Mathématiciens : arbre associé aux sujets dans la catégorisation contrainte (avec ajout d'objets fictifs)

## II.4 Comparaison des arbres des deux catégorisations

La comparaison des structures d'arbres pour les situations montre un point commun, à savoir la séparation entre les situations réelles et les situations stochastiques ; néanmoins cette distinction est sensiblement atténuée dans la phase 2. Cependant la comparaison révèle également plusieurs différences dont les principales sont les suivantes. Pour les situations stochastiques le type d'issue « symétrique » ou « non symétrique » intervient beaucoup moins. Pour les situations réelles, les mathématiciens ne font plus la distinction entre « avec » ou « sans » implication ; pour ces situations encore, les situations A (Ami) et B (Loto) sont plus proches des situations stochastiques.

Ces différences dans les structures d'arbre montrent des changements notables d'une phase à l'autre. On peut supposer que le fait de rendre explicite la notion de hasard conduit le sujet à activer des connaissances différentes ce qui entraîne un changement de point de vue sur les situations considérées. Ceci renvoie à l'aspect dynamique de la catégorisation et à la flexibilité des représentations du sujet.

### III. Interprétations : quelques représentations du hasard

Dans la situation de catégorisation avec contrainte, les résultats ont été complétés et éclairés notamment par une analyse des patrons de réponses individuels, s'appuyant sur les justifications données par les sujets. Cette analyse, non présentée ici, permet de dégager cinq interprétations dominantes: deux interprétations globales rendant compte de l'ensemble des 16 réponses et trois interprétations locales rendant compte d'une partie des réponses

La première interprétation globale dominante est « il y a hasard quand il y a un raisonnement probabiliste » ; ceci revient à considérer que le hasard intervient toujours. Cette interprétation est majoritaire, surtout chez les mathématiciens, ce qui n'est pas surprenant. Cependant, on observe (surtout chez les mathématiciens) une distinction entre deux types de hasard, selon les situations: un hasard mathématique lié aux événements pour lesquels on peut calculer une probabilité, comme c'est le cas pour les situations stochastiques (conception objectiviste classique) ; un hasard par ignorance lié aux événements pour lesquels la probabilité est calculable mais difficilement, comme c'est le cas pour les situations réelles. Donc il y a toujours du hasard mais les situations stochastiques et réelles sont opposées selon la nature du hasard.

La deuxième interprétation globale oppose également les situations stochastiques et les situations réelles mais selon la présence ou l'absence du hasard. Les sujets ont un point de vue déterministe sur les événements réels (donc « le hasard n'intervient pas ») ; par contre, on peut calculer une probabilité pour les items stochastiques (donc « le hasard intervient »). C'est chez les collégiens qu'on rencontre le plus cette interprétation ; il est intéressant de noter qu'avant tout enseignement des probabilités les collégiens repèrent les situations facilement probabilisables.

Les trois interprétations locales, dont chacune rend compte d'une partie des réponses sont les suivantes. (1) Un événement relève typiquement du hasard quand toutes les issues ont la même probabilité, ce qui se traduit par la réponse que le hasard intervient au moins pour les quatre situations stochastiques à issues symétriques ; ce type d'interprétation est à rapprocher du « biais d'équiprobabilité » (Lecoutre, 1992) selon lequel des événements aléatoires sont considérés comme équiprobables *par nature*. (2) Le degré d'intervention du hasard est lié à la valeur des probabilités associées aux issues, le hasard intervenant d'autant plus que ces valeurs sont faibles. (3) Enfin, un phénomène relève du hasard quand on ne peut pas faire de prédiction soit parce qu'on ne peut pas identifier de facteur causal, soit parce qu'on n'a aucun contrôle sur la situation.

On notera encore que les enseignants-chercheurs en psychologie se différencient par le fait que 3 sujets disent que le hasard n'intervient jamais ; c'est une interprétation qui n'est pas rencontrée dans les deux autres groupes. De plus, il faut noter que la justification suivante a été également trouvée uniquement chez les psychologues : « dès que l'on peut calculer une probabilité il n'y a plus de

hasard » ; on peut se demander dans quelle mesure ceci n'est pas lié à leur pratique des tests de signification (*cf.* notamment Lecoutre, Lecoutre et Poitevineau, 2001), où le calcul d'un seuil est utilisé pour rejeter l'hypothèse du hasard.

#### **IV. Conclusion : intérêt et limites de la méthode**

La méthode des arbres additifs s'est révélée utile. Elle permet une structuration plus fine que celle que l'on aurait pu obtenir par une analyse des proximités (Cox et Cox, 2001). Ainsi ici elle nous a notamment permis de mettre en évidence la flexibilité des représentations suivant l'explicitation ou non de la notion de hasard.

Cependant les arbres obtenus sont, en quelque sorte, des arbres « moyens » calculés à partir des données agrégées dans chaque groupe de sujets, ce qui limite leur interprétation ; néanmoins celle-ci peut être affinée, dans la catégorisation contrainte, en examinant les patrons de réponses individuels. La construction d'arbres individuels nécessiterait une autre procédure de recueil des données pour pouvoir obtenir une matrice de dissimilarités pour chaque sujet. Sans que cela soit impossible, une telle procédure serait irréaliste et en outre très lourde dès que le nombre d'objets en jeu est un tant soit peu élevé (ici, un sujet devrait effectuer  $16 \times 15 / 2 = 120$  comparaisons par paires pour apprécier toutes les dissimilarités entre les 16 situations).

Enfin, comme pour toute méthode de statistique descriptive, les résultats obtenus ne sont, en principe, applicables qu'aux sujets examinés. La généralisation à une population plus vaste pourrait s'appuyer sur des études de simulation, mais ce travail reste à faire et la généralisation est sans doute problématique.

## Références

- Barsalou, L. W. (1987) The instability of graded structure : Implications for the nature of concepts, In U. Neisser (éd.), *Concepts and conceptual development : Ecological and intellectual factors in categorization*, 101-141, Cambridge, Cambridge University Press.
- Barsalou, L. W. (1993) Flexibility, structure, and linguistic vagary in concepts : manifestations of a compositional system of perceptual symbols, In A. F. Collins, S. E. Gathercole, M. A. Conway & P. E. Morris (éds), *Theories of memory*, Hillsdale, N. J., Erlbaum.
- Barthélémy, J. -P. & Guénoche, A. (1991) *Trees and Proximity Representations*, Wiley.
- Blaye, A. & Bonthoux, F. (2001) Thematic and taxonomic relation in preschoolers : the development of flexibility in categorization choices, *British Journal of Developmental Psychology*, **19**, 395-412.
- Cox, T. F. & Cox, M. A. A. (2001) *Multidimensional Scaling. 2<sup>nd</sup> edition*, Chapman & Hall.
- Guénoche, A. & Garreta, H. (2001) Can we have confidence in a tree representation ? Proceedings of JOBIM'2000, *Lecture Notes in Computer Sciences*, vol. 2066, 43-53.
- Konold, C., Falk, R., Lipson, A., Lohmeier, J., Pottatsek, A. & Well, A. (1991) Novices views on randomness, Paper presented at the 13<sup>th</sup> Annual Meeting on the International Group for the Psychology of Mathematics Education, Blacksburg, VA.
- Lecoutre, B., Lecoutre, M.-P. & Poitevineau, J. (2001) Uses, abuses and misuses of significance tests in the scientific community : won't the Bayesian choice be unavoidable? *International Statistical Review*, **69**, 399-417.
- Lecoutre, M.-P. (1992) Cognitive models and problem spaces in « purely random » situations, *Educational Studies in Mathematics*, **23**, 557-568.
- Medin, D. L., Lynch, E. B. & Coley, J. D. (1997) Categorization and reasoning among tree experts : do all roads led to Rome ? *Cognitive Psychology*, **32**, 49-96.
- Nisbett, R. E., Krantz, D. H., Jepson, C. & Kunda, Z. (1983) The use of statistical heuristics in everyday inductive reasoning, *Psychological Review*, **90**, 339-363.
- Robinson, D. R. & Foulds, L. R. (1981) Comparison of phylogenetic trees, *Mathematical Biosciences*, **53**, 131-147.
- Sattath, S. & Tversky, A. (1977) Additive similarity trees, *Psychometrika*, **42**, 319-345.

# Groupements d'odeurs ou catégories ? Questions de méthodologie

*Sophie David*<sup>(1)</sup>

*Catherine Rouby*<sup>(2)1</sup>

(1)(CNRS UMR 7114 - Modyco et Université Paris 10)

(2)(CNRS UMR 5020 et Université Claude Bernard Lyon 1)

## Introduction

La catégorisation des odeurs est considérée comme le domaine emblématique du désaccord et de la subjectivité par excellence, où ni le discours scientifique ni le bon sens ne proposent de catégories universellement ou même généralement acceptées. Les langues elles-mêmes, tout au moins les langues indo-européennes, sont « défailtantes » à proposer des termes spécifiques, à partir desquels on pourrait inférer une organisation catégorielle. Depuis quelques années, nous avons initié un travail de défrichage des aspects psychologiques et linguistiques de la catégorisation des odeurs (Dubois & Rouby 1997). Ces travaux confirment l'absence de grille universelle de représentation des odeurs (Rouby & Sicard 1997 ; Chastrette 2002) et, dans le même temps, la cohérence des sujets chargés de les catégoriser : ils sont au moins d'accord sur l'absence de norme linguistique et de connaissances partagées (David *et al.* 1997 ; David 1997 ; David *et al.* 2000). L'examen d'autres langues (Boisson 1997) et d'autres cultures (Schaal *et al.* 1998) nous a permis de montrer que le jugement hédonique est le mieux représenté dans les langues, et que sa valeur négative – les mauvaises odeurs – qui partitionne régulièrement l'espace olfactif, est la plus stable entre les sujets et entre les cultures. L'étude de ce jugement a également été abordée sous l'angle neurophysiologique, et nous avons proposé des arguments pour un traitement différentiel des odeurs désagréables (Rouby & Bensafi 2002, Bensafi *et al.* 2001, 2002a, b et c).

La littérature psychophysique en olfaction ne fait jamais allusion aux concepts de catégorisation, mais les méthodes d'analyse de proximité (*multidimensional*

---

<sup>1</sup>Nous remercions vivement Jacques Poitevineau et Danièle Dubois pour leurs relectures attentives.

*scaling*) ont été largement utilisées pour explorer l'organisation de l'espace des odeurs (Schiffman *et al.* 1977 ; Berglund *et al.* 1973 ; Carrasco & Ridout 1993). Le travail consiste le plus souvent à estimer des ressemblances perceptives entre odorants<sup>2</sup>, le plus souvent en demandant aux sujets de chiffrer la ressemblance pour chaque paire d'odorants, et à extraire ensuite de cette matrice de distances perceptives les dimensions qui les représentent au mieux. Plusieurs études ont aussi essayé de corrélérer ces dimensions à des paramètres physico-chimiques des corps purs ainsi évalués (Schiffman *et al.* 1974).

L'analyse factorielle des correspondances et d'autres méthodes multidimensionnelles sont aussi très utilisées en analyse sensorielle, où des échelles de jugement et des descripteurs servent à composer un espace dans lequel différents objets, le plus souvent des aliments ou des boissons, se regroupent ou s'opposent (Issanchou *et al.* 1989 ; Aubry *et al.* 1999 ; Urdapilletta *et al.* 2001).

Ces différentes méthodes permettent de mettre en évidence des groupements d'odeurs ; la question est alors de savoir s'ils constituent des catégories, renvoyant à une organisation cognitive stable, qui resterait valable à travers les différents contextes. En ce qui concerne l'organisation de ces catégories entre elles, toutes les études citées constatent que, si elle existe, elle est très peu hiérarchisée.

Nous avons repris ces questions en cherchant d'abord à identifier quels étaient les principes organisateurs des catégories, non plus à partir de jugements sur échelles ou de comparaisons par paires, mais à partir de tâches de catégorisation. Nous avons ainsi proposé d'étendre aux odorants la procédure de groupement utilisée en vision et en audition, qui consiste à ne pas imposer aux sujets un ordre strict de présentation des stimuli, ni des descripteurs *a priori*, et à leur demander de constituer progressivement des groupes d'odeurs selon leurs ressemblances (Godinot 1994 et 1999 ; Dubois *et al.* 1997 ; Dubois 2000). Ces tâches de catégorisation comportent en outre une phase de verbalisation : les sujets doivent fournir des descriptions identifiant les groupements formés, s'exprimer sur les critères qu'ils pensent avoir mobilisés, identifier les odorants testés, etc.

Si, pour traiter les résultats obtenus, l'utilisation de méthodes mathématiques objectivant des dimensions paraît dans un premier temps appropriée et intéressante, elle pose aussi un problème important, car on applique sur les données un modèle dont on ne sait *a priori* s'il est pertinent pour les objets étudiés. La prudence méthodologique s'impose. Elle demande de connaître précisément les choix, voire les biais, introduits dès le départ dans la notation des données ; elle demande de connaître précisément le modèle sous-jacent ; elle demande une discussion précise des critères permettant d'évaluer les résultats. C'est seulement à ces conditions que l'on pourra alors confirmer ou infirmer les hypothèses testées.

Nous nous intéresserons ici à des méthodes plus souvent utilisées en phylogénie, des méthodes fondées sur les arbres de similarité additifs. Et en particulier à celle

---

<sup>2</sup> Nous appelons « odorant » le stimulus sensoriel qui a pour fonction dans un contexte expérimental de représenter une odeur. Dans les expériences présentées ici, il s'agira de dilutions d'arômes et de corps purs. Sur la distinction entre odeur et odorant, voir Dubois et Rouby (2002).

qui repose sur l'évaluation de scores d'association de paires d'objets, proposée par Sattath et Tversky (1977)<sup>3</sup>. Elle a été implémentée par Guénoche et Garreta (2001). Son intérêt est de proposer une organisation topologique des différents groupements émergeant à partir des données individuelles (ce qui n'est pas donné dans les méthodes d'analyse de proximité). De l'arbre construit, on peut dériver immédiatement (par coupures de certaines branches) des classes, qui, dans certaines conditions, peuvent être analysées comme des catégories. Le logiciel fournit en outre un certain nombre de critères permettant d'évaluer la qualité de la représentation fournie : les critères métriques classiques (tels que le stress de Kruskal, l'écart moyen, etc.), mais aussi des critères évaluant précisément la topologie construite.

Par la suite, on fera référence à cette méthode sous le nom de Addtree.

Nous souhaitons ainsi expliciter :

- 1) les choix possibles au cours des étapes du traitement, et leurs conséquences ;
- 2) les critères permettant de choisir entre des traitements différents des mêmes données ;
- 3) les critères qui permettent d'évaluer la représentation obtenue.

Pour apprécier l'intérêt de l'outil utilisé (Addtree), nous nous proposons de mettre en place une méthode d'analyse, qui croise les différents résultats produits par le logiciel avec les différents types de données sources (analyse des partitions, analyse des verbalisations). Si on montre comment l'on peut objectiver le(s) fondement(s) des organisations catégorielles étudiées, on aura du même coup montré la valeur heuristique de ce type d'outil.

Si notre objectif est avant tout d'ordre méthodologique, la discussion sera menée à partir des résultats des expériences sur les odeurs. Deux aspects seront plus spécifiquement abordés :

- (i) la caractérisation du jugement hédonique ;
- (ii) le rôle des pratiques et des activités dans la catégorisation.

Ce faisant, nous serons amenées à mentionner d'autres propriétés, telles que celles d'intensité, de familiarité et de typicité.

Les hypothèses qui ont présidé à la conception des différentes expériences feront l'objet des première et deuxième parties. Nous nous attacherons dans la troisième partie à la présentation des données et à leur traitement, en faisant une brève présentation du logiciel utilisé. La quatrième partie sera consacrée à l'analyse des résultats : on examinera tout d'abord les résultats issus de traitements opérés avec des valeurs de paramètres différents ; on procèdera ensuite à une analyse globale, puis à des analyses plus spécifiques : elles porteront sur les classes auxquelles on peut accorder ou non une certaine fiabilité. A cette occasion, on développera la méthode d'analyse présentée succinctement ci-dessus.

---

<sup>3</sup> Nous laissons de côté la méthode NJ (neighbor-joining method) proposée par Saitou & Nei (1987), qui produit sur les données des résultats légèrement inférieurs, et pour laquelle nous disposons de moins de critères implémentés.

## 1. Hypothèses sous-jacentes aux différentes expériences

Les données présentées ici sont issues de tâches de catégorisation libre, conçues à partir du cadre de recherche proposé par Dubois (1991, 1997) : laisser les sujets « libres » de catégoriser (rassembler et distinguer) les objets testés selon leurs propres paramètres évite la projection de modèles *a priori* ; le travail consiste alors à identifier les fondements des catégories produites. Dans un deuxième temps, le modèle que l'on aura élaboré peut être confronté à d'autres, notamment à ceux qui ont été proposés pour la vision. Pour mener à bien cette tâche, le matériel testé doit être « écologiquement valide » (Dubois 1993), plus précisément ici les stimuli olfactifs doivent être reconnus par les sujets comme des odeurs familières.

Le domaine olfactif présente par ailleurs une propriété remarquable. Son expression sur le plan linguistique dans la plupart des langues (Boisson 1997), et notamment en français (David 2002), a la particularité de ne présenter que peu de « termes spécifiques ». Excepté quelques termes majoritairement utilisés tels que *odeur, parfum, etc.*, on ne dispose pas de termes dénotant une odeur particulière<sup>4</sup>. Ce point fait fortement contraste avec le domaine visuel et en particulier celui des couleurs, pour lequel on a des termes spécifiques tels que *bleu, rouge, jaune, etc.*, et aussi des termes référant à des objets porteurs de couleurs particulières tels que *azur, cerise, brique, etc.* Rien de tel pour le français, ni du côté de la première série, ni du côté de la seconde<sup>5</sup>. C'est pourquoi il est particulièrement intéressant de combiner la tâche de catégorisation à une tâche de verbalisation. On peut ainsi étudier l'influence de cette situation linguistique particulière sur les processus de catégorisation<sup>6</sup>.

Différentes expériences ont été menées. Elles ont été construites sur les trois variables suivantes :

- la présence d'un « nom d'odeur » ;
- la présence du stimulus olfactif ;
- la consigne perceptive ou émotionnelle.

Nous entendons par « nom d'odeur » le « nom attendu » (*veridical label*), *i.e.* le « nom » par lequel une odeur est identifiée. Cette première variable, fournir ou

---

<sup>4</sup> *Remugle, miasme, fraîcheur* (terme régional du sud-ouest) qui dénotent des odeurs particulières (respectivement le renfermé, un gaz pestilentiel dégagé par les déchets et une odeur de poisson) restent des cas isolés.

<sup>5</sup> Comme l'ont noté Corbin et Temple (1994), il n'y a aucun processus morphologique permettant de dériver un nom d'odeur à partir d'un nom dénotant un « objet » ayant une odeur typique (par exemple, le thym, le musc). Les noms associés aux odeurs restent des noms dénotant des « objets ». Ce point fait contraste avec d'autres procédés linguistiques qui « voient » les aspects visuels. On l'a dit pour les couleurs, cela est aussi vrai de la forme. Par exemple une pomme, une poire, dénotent des fruits mais aussi des objets ayant la forme d'une pomme, d'une poire, etc. Par exemple, la pomme d'une canne, une poire à lavement.

<sup>6</sup> Distel et Hudson (2001) ont montré que le jugement d'intensité variait selon que l'odorant était étiqueté ou pas d'un nom.



pas aux sujets le « nom attendu » se justifie par un ensemble de faits. Tout d'abord par ce que nous avons exposé succinctement ci-dessus : la situation particulière pour beaucoup de langues de ne pas disposer de termes particuliers pour désigner des odeurs. Ensuite, par le fait maintes fois noté dans la littérature : la difficulté pour des sujets d'identifier des odeurs, y compris des odeurs familières (Engen 1987). Les sujets sont ou se considèrent particulièrement démunis dans ce type de tâches<sup>7</sup>. Enfin, par le fait que les odorants sont généralement étiquetés par l'expérimentateur d'un nom<sup>8</sup>, le « nom attendu », sans que soit justement interrogé le rapport entre l'odorant, l'odeur et la dénomination utilisée<sup>9</sup>. Pour explorer dans quelle mesure la présence préalable d'un nom contraint la catégorisation, nous avons utilisé dans certaines expériences des flacons odorants comportant le nom attendu, et dans les autres des flacons ne comportant pas de nom.

La seconde variable que nous avons introduite, présence *vs* absence de stimulus, est symétrique de la première. Des expériences de catégorisation à partir des noms seuls, donc sans stimuli olfactifs, utilisent des étiquettes. Sur chacune d'entre elles est inscrit « le nom attendu ». Pour effectuer la tâche, les sujets disposent uniquement de leur mémoire sémantique des odeurs.

La troisième variable, perception *vs* émotion, se justifie à partir de la spécificité des odeurs, qui est de provoquer plaisir/déplaisir et de susciter des souvenirs. Leur forte capacité d'évocation constitue alors un terrain propice pour étudier les rapports entre perception et émotion. Nous avons demandé à une partie des sujets de classer les odorants en fonction de leurs ressemblances perceptives ; les autres sujets devaient les classer en fonction des émotions que ces odeurs avaient suscitées. Ayant développé par ailleurs l'hypothèse que la catégorisation spontanée des odeurs est émotionnelle (Bensafi *et al.*, 2001 ; 2002a, b, c), nous attendions des différences entre les catégorisations qui ont été amorcées par des noms et celles qui ne l'ont pas été.

## 2. Présentation des expériences

Dans cette section, nous présentons successivement le matériel expérimental et les consignes qui ont été données aux sujets.

### 2.1. Le matériel expérimental

Nous disposons de trois matériels expérimentaux :

---

<sup>7</sup> Ce que nous avons aussi observé lors de nos premières expériences, cf. ci-dessous l'expérience réalisée en 1998.

<sup>8</sup> Dans des tâches psycho-sensorielles classiques d'identification.

<sup>9</sup> La situation est complexe, car les odeurs n'ont pas de noms/*names*. Dans les faits, on utilise le nom attendu (qui est par ailleurs un *name*) comme une désignation. Sur la distinction dénomination (*name*) et désignation, voir Kleiber (1984).

- 1) 16 flacons contenant chacun un odorant familier, chaque flacon est étiqueté avec un nom ;
- 2) 16 flacons contenant chacun l'un de ces 16 mêmes odorants étiquetés uniquement avec un numéro ;
- 3) 16 étiquettes portant chacune l'un des noms utilisés en 1).

Numéro	Nom	Numéro	Nom	Numéro	Nom	Numéro	Nom
1	vanille	5	poisson	9	orange	13	anis
2	lavande	6	violette	10	pomme	14	mûre
3	eucalyptus	7	ail	11	cannelle	15	chewing-gum
4	fuel	8	herbe	12	citron	16	menthe

Tableau 1 : identification des odorants (numéro + « nom »)

Ce matériel expérimental permet d'introduire les deux variables suivantes :

- présence ou absence de stimulus olfactif ;
- présence ou absence du nom attendu dans les expériences avec stimulus olfactif.

On désignera les expériences de la manière suivante :

Avec stimulus olfactif		Sans stimulus olfactif
<b>anoms</b> ('avec les « noms »')	<b>snoms</b> ('sans les « noms »')	<b>étiq</b> ('étiquette')
identification des odorants par un « nom »	identification des odorants par un numéro	

Tableau 2 : identification des expériences

## 2.2. La tâche et les consignes

A partir de ces différents matériels, nous avons demandé aux sujets d'effectuer une tâche de catégorisation libre : il s'agissait de mettre ensemble les « matériels » qui se ressemblent, soit sur la base d'une ressemblance perceptive, soit sur la base d'une ressemblance du point de vue émotionnel (cf. § 1). Pour l'expérience sans stimulus olfactif (expérience « étiquettes »), il s'agissait d'appliquer la même consigne à partir de l'odeur évoquée par le nom indiqué sur l'étiquette.

Une fois la catégorisation achevée, les sujets remplissaient une fiche de résultats, dans laquelle ils indiquaient les odorants/odeurs qui avaient été regroupés (soit par leur nom, soit par leur numéro) et proposaient une description appropriée pour chaque groupe qui avait été réalisé.

Soient les consignes suivantes :

*a* Expérience « *anoms* » / « *snoms* » + *ressemblance perceptive*

Voici un ensemble de 16 odeurs. Nous vous demandons de les regrouper selon leurs ressemblances et leurs différences.

*b* Expérience « *anoms* » / « *snoms* » + *ressemblance du point de vue des émotions suscitées*

Voici un ensemble de 16 odeurs. Nous vous demandons de regrouper ces odeurs en fonction des émotions qu'elles vous suscitent.

*c* Expérience « *étiq* » + *ressemblance perceptive*

Voici un ensemble de 16 étiquettes. Sur chaque étiquette est indiqué un nom évoquant une odeur. Nous vous demandons de regrouper ces odeurs selon leurs ressemblances et leurs différences.

*d* Expérience « *étiq* » + *ressemblance du point de vue des émotions suscitées*

Voici un ensemble de 16 étiquettes. Sur chaque étiquette est indiqué un nom évoquant une odeur. Nous vous demandons de regrouper ces odeurs en fonction des émotions qu'elles vous suscitent<sup>10</sup>.

On récapitulera les différentes expériences de la manière suivante<sup>11</sup> :

Consigne	Avec stimulus olfactif		Sans stimulus olfactif	
	avec les noms	sans les noms		
<b>Emotion</b>	anoms-é-00	snoms-é-00		étiq-é-00
<b>Perception</b>	anoms-p-01	snoms-p-98 snoms-p-01	étiq-p-01	

Tableau 3 : récapitulatif des différentes expériences

<sup>10</sup> Les consignes a et b finissaient par « Après avoir senti un flacon, prenez soin de le refermer. Vous pouvez faire autant de groupes que vous le souhaitez. Un groupe peut comprendre autant de flacons que vous le voulez. Vous disposez d'autant de temps que vous le souhaitez. Quand vous avez terminé, merci de bien vouloir remplir la feuille jointe. »

Les consignes c et d finissaient par « Vous pouvez faire autant de groupes que vous le souhaitez. Un groupe peut comprendre autant d'étiquettes que vous le voulez. Vous disposez d'autant de temps que vous le souhaitez. Quand vous avez terminé, merci de bien vouloir remplir la feuille jointe. »

<sup>11</sup> Lors des expériences 2001, les sujets devaient indiquer pour chaque classe réalisée quel(le) était l'odeur/l'odorant le plus typique.

Pour l'expérience « snoms-p-98 », on peut noter que :

- les sujets ont eu globalement plus de temps que les sujets qui ont passé les expériences 2000 et 2001 ;

- en plus de la catégorisation, les sujets avaient une tâche d'identification (ils devaient dire pour chaque odorant (senti de nouveau) de quelle odeur il s'agissait) et une tâche d'explicitation (ils devaient indiquer, pour les deux tâches précédentes, quels étaient les critères qu'ils avaient mobilisés) ; leurs commentaires ont été enregistrés à l'aide d'un magnétophone et retranscrits ; on dispose ainsi de beaucoup plus d'informations sur le processus temporel des groupements (hésitations, retours en arrière, justifications, etc.) ;

- pour la tâche de catégorisation, 4 ordres de passation des odorants ont été prévus et réalisés.

Nous n'analyserons pas en détail les 2<sup>ème</sup> et 3<sup>ème</sup> parties de l'expérience. Nous nous en servirons le cas échéant pour justifier certaines de nos hypothèses.

- la consigne « perception » est identifiée par « p » ; la consigne « émotion » par « é » ;
- ces expériences ont été réalisées en 1998, 2000 et 2001 : les deux chiffres notés en dernière position y font référence.

### 2.3. Les sujets

Les expériences réalisées en 2000 et 2001 sont homogènes du point de vue des sujets : des étudiants de 25-30 ans. Pour l'expérience réalisée en 1998, les sujets avaient entre 30 et 50 ans. Deux sujets avaient plus de 60 ans.

Expérience	Nb sujets	Expérience	Nb sujets	Expérience	Nb sujets
anoms-é-00	19	snoms-p-98	40	étiq-é-00	21
anoms-p-01	19	snoms-é-00	17	étiq-p-01	20
		snoms-p-01	24		

Tableau 4 : nombre de sujets par expérience

## 3. Les données et leurs traitements

Le traitement des données est effectué par deux modules<sup>12</sup> : le premier construit une matrice des distances entre les objets ; le second cherche à représenter ces distances par des arbres d'un type particulier, des arbres de similarité additifs, en évaluant des scores de paires d'objets (Sattath et Tversky 1977). Il fournit en outre des critères permettant d'apprécier la qualité de la représentation produite.

Dans cette section, nous présentons les différentes étapes du traitement informatique en explicitant les problèmes, les choix et les paramètres qui y sont reliés.

### 3.1. Les données

Les données sont constituées par les résultats de chacune des expériences. Pour chaque sujet, on dispose de l'ensemble des classes<sup>13</sup> qu'il a effectuées. Le premier problème qui se pose est le suivant : quand un sujet effectue un classement hiérarchisé, doit-on tenir compte des proximités introduites par la hiérarchie ? Et comment ?

---

<sup>12</sup> Mis au point par Guénoche & Garreta (2001) et disponibles à l'adresse suivante : <http://www.fas.umontreal.ca/biol/legendre>. Ils ont été réaménagés par J. Poitevineau.

<sup>13</sup> Ces classes forment une partition des objets, *i.e.* un ensemble non vide de classes, telles que leur intersection est vide et que leur union produit l'ensemble des objets. Un odorant est donc forcément dans une classe et une seule.

Soit par exemple la situation où un sujet a proposé les classes suivantes :

Ex. :	classe 1 :	les odeurs désagréables
	classe 2 :	les odeurs agréables comprenant elle-même deux sous-classes :
		classe 2' : les odeurs de fruits
		classe 2'' : les odeurs de fleurs
	classe 3 :	les odeurs fortes

Trois classes ont été proposées, dont l'une, la classe 2, est constituée de deux sous-classes (2' et 2'').

On pourrait vouloir indiquer que les odeurs de fruits et les odeurs de fleurs sont plus proches que les odeurs de fruits ne le sont des odeurs désagréables ou que les odeurs de fleurs ne le sont des odeurs fortes. La difficulté est alors de trouver la mesure de cette proximité.

Il n'y a pas de solution simple à ce problème. Par ailleurs, l'algorithme implémenté dans l'outil utilisé ne permet pas de prendre en charge ce problème. On a donc dû choisir au départ le niveau de traitement. Ayant observé que la plupart des sujets ne construisaient que peu de hiérarchie, nous avons décidé de traiter le niveau le plus bas (dans notre exemple, celui des « fruits » et « des fleurs »). Le fichier de données est ainsi constitué des ensembles de classes de niveau le plus bas, réalisés par chaque sujet.

### 3.2. Les paramètres de la construction de la matrice des distances

Avant de lancer la construction de la matrice des distances, on doit fixer deux paramètres :

- le paramètre « objet/sujet » ;
- le paramètre « prise en compte du nombre de classes ».

#### *Le paramètre « objet/sujet »*

On peut choisir d'obtenir une représentation des objets qui ont été classés et/ou une représentation des sujets en fonction de leur classement. Dans le cadre de cet article, nous ne traiterons que des objets (*i.e.* les odorants).

#### *Le paramètre « prise en compte du nombre de classes »*

On considère que, pour un sujet donné, la distance entre deux objets se trouvant dans la même classe est égale à 0 et que la distance entre deux objets se trouvant dans deux classes est égale à 1<sup>14</sup>. Soit la situation suivante<sup>15</sup>, un sujet a fait autant

---

<sup>14</sup> Transformer le problème de l'appartenance à une classe (réponse binaire) en une distance constitue un point problématique. En effet :

- distance = 0 : tous les objets d'une même classe sont équidistants. Ce traitement n'est pas forcément justifié. Il contredit même l'hypothèse d'une organisation catégorielle prototypique (cf. Rosch *et al.* 1976 ; Rosch 1978 ; Dubois 1991) ;

- distance = 1 : deux objets, a et b, appartenant respectivement aux classes A et B sont équidistants d'un objet c appartenant à la classe C. Or a peut être plus proche de c que ne l'est b.

de classes que d'objets. Dans ce cas, sa contribution à la matrice générale des distances est maximale (tous les objets sont tous distants les uns des autres). Ainsi, plus un sujet fait de classes, plus il contribue à la matrice générale des distances. Tenir compte du nombre de classes revient à accorder aux sujets le même poids : on impose alors que la contribution de chacun des sujets soit une constante ; la distance entre les objets est pondérée en fonction du nombre de classes. Dans le cadre de ces expériences, en l'absence de justifications permettant d'opter *a priori* pour l'un ou l'autre, nous avons doublé les traitements. Les traitements où l'on aura tenu compte du nombre de classes seront identifiés par « + », ceux pour lesquels on n'aura pas tenu compte du nombre de classes seront identifiés par « - ». Les résultats feront l'objet de la discussion du § 4.1.

### 3.3. La construction de la matrice des distances entre les objets

Une fois ces paramètres fixés, le fichier de données est traité par un premier module qui produit la matrice des distances entre les objets pour tous les sujets. La distance entre deux objets a et b,  $D(a, b)$ , est le pourcentage de sujets qui ont placé a et b dans des classes différentes, multiplié par 100 (Guénoche 1999). Par exemple, si sur 10 sujets, 7 ont placé a et b dans des classes différentes,  $D(a, b) = 70$ . La distance peut donc varier entre 0 (identité) et 100 maximum.

	1 - Vanille	2 - Lavande	3 - Eucalyptus	4 - Fuel	5 - Poisson	6 - Violette
1 - Vanille	0,00	84,21	94,73	100,00	100,00	68,42
2 - Lavande	84,21	0,00	89,47	94,73	100,00	47,36
3 - Eucalyptus	94,73	89,47	0,00	100,00	100,00	89,47
4 - Fuel	100,00	94,73	100,00	0,00	36,84	100,00
5 - Poisson	100,00	100,00	100,00	36,84	0,00	100,00
6 - Violette	68,42	47,36	89,47	100,00	100,00	0,00

Tableau 5 : extrait de la matrice des distances pour « anoms-é-00 »

Dans « anoms-é-00 »<sup>16</sup>, la vanille n'a jamais été regroupée avec le fioul ou le poisson (distance de 100), elle est plus proche de la violette que de la lavande (68,42 vs 84,21).

---

Il n'y a pas de solution idéale, en particulier dans la situation qui est la nôtre : on cherche à appliquer des modèles sur des données dont on ne sait *a priori* quel est leur modèle d'organisation. Le choix de telle ou telle méthode introduit *de facto* des décalages, des biais, qu'il importe de garder à l'esprit.

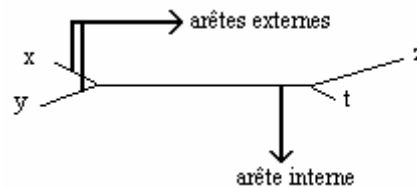
<sup>15</sup> Si un sujet fait une seule classe, sa contribution à la matrice générale des distances est égale à 0 (aucun objet n'est distant d'aucun autre). Un tel sujet n'influencerait en aucune manière les résultats.

<sup>16</sup> Pour simplifier la présentation, nous notons les distances avec uniquement deux décimales, et ce, pour les 6 premiers odorants.

### 3.4. Le traitement selon la méthode Addtree et les résultats

Le fichier contenant la matrice des distances constitue le fichier de données pour le second module. A l'issue du traitement, on obtient trois types de résultats :

- (1) une représentation en termes d'arbres de similarité additifs. Les feuilles de l'arbre sont les objets traités. Une arête est un chemin entre deux nœuds (point de rencontre). Une arête dite externe est une arête qui « raccorde » un objet (une feuille) au reste de l'arbre. Toutes les autres arêtes sont dites internes. Les arêtes sont valuées : elles ont une certaine longueur, qui indique une distance.



L'arbre construit forme une topologie. La représentation graphique des résultats peut être construite à l'aide du logiciel Treeview (cf. exemple en annexe) ;

- (2) les longueurs des arêtes internes et externes, de même que leur longueur moyenne ;
- (3) deux séries de critères, métrique et topologique, qui permettent d'évaluer la qualité de la représentation obtenue. Nous renvoyons à Guénoche (1999) pour la définition formelle de chacun d'entre eux.

Dans cet article, nous nous intéresserons à quatre d'entre eux, un critère métrique et trois critères topologiques :

**Critère métrique** Stress de Kruskal

**Critères topologiques** Taux de quadruplets bien représentés  
Taux de quadruplets mal représentés impliquant un objet  
Taux de quadruplets supportant chaque arête interne

- Le stress de Kruskal mesure l'écart entre la matrice initiale des distances (les données des sujets) et la matrice des distances, construite par la méthode Addtree (matrice de la représentation produite). La représentation produite est parfaite quand le stress est égal à 0 ;
- Le taux de quadruplets bien représentés mesure le pourcentage de quadruplets qui présentent la même topologie dans la matrice initiale des distances (les données des sujets) et la matrice des distances de la représentation obtenue. Plus le nombre est élevé, plus la représentation fournie peut être considérée comme « fiable »<sup>17</sup> ;

---

<sup>17</sup> « Fiable » est ici entendu dans un sens non technique. Il renvoie au fait que la représentation (ou une ou plusieurs de ses parties) est associée à des critères ayant une valeur élevée, et que de ce fait elle mérite discussion.

- Le taux de quadruplets mal représentés impliquant un objet mesure le pourcentage de quadruplets qui ne présentent pas la même topologie dans la matrice initiale des distances (les données des sujets) et la matrice des distances de la représentation obtenue, et ce pour un objet donné ;
- Le taux de quadruplets supportant chaque arête interne indique la qualité de l'arête interne. Plus le nombre est élevé, plus l'arête interne peut être considérée comme « fiable ».

## 4. Analyses

Pour analyser les données, nous prendrons en compte deux types de ressources :

- (i) les données sources, elles-mêmes de deux natures différentes : les partitions et les verbalisations ;
- (ii) les résultats produits par le traitement Addtree : la topologie construite, les critères métriques et les critères topologiques.

### (i) Les données sources

Pour les partitions, on examinera :

- les objets associés, *i.e.* les objets qui n'appartiennent pas à la classe considérée mais qui y ont été associés par des sujets<sup>18</sup> ;
- les objets formant une classe unique ;
- les sous-classes, *i.e.* les classes impliquant un sous-ensemble des objets d'une classe considérée.

Pour les verbalisations, on isolera dans un premier temps les descriptions des sujets qui ont constitué une classe comprenant au moins les objets constituant une classe pour Addtree. Dans un deuxième temps, on essaiera de repérer des régularités linguistiques, tant sur la catégorie des formes employées (syntagme nominal, adjectival, etc.) que sur leur type linguistique (événement, propriété, etc.). Par la suite, on examinera les domaines référentiels dénotés par les expressions utilisées.

---

<sup>18</sup> Par exemple : deux sujets ont constitué respectivement « fuel, poisson, ail, herbe » et « fuel, poisson, ail, cannelle ». On considérera qu'il y a deux objets associés à « fuel, poisson, ail », herbe et cannelle.



(ii) Les résultats produits par la méthode Addtree  
 Nous nous appuyerons essentiellement sur :

- la topologie en tant que telle ;
- la longueur des arêtes externes ;
- le stress de Kruskal ;
- le taux de quadruplets bien représentés ;
- le taux de quadruplets supportant chaque arête interne ;
- le taux de quadruplets mal représentés impliquant tel objet.

Dans la suite de cet article, nous discuterons :

- des résultats en fonction du paramètre « prise en compte du nombre de classes » ;
- de la méthode pour isoler des « groupements fiables » ;
- de la structuration des « groupements fiables » ;
- du statut des objets appartenant à des « groupements fiables » ;
- de l'identification des objets spécifiques ;
- des objets appartenant à des « groupements peu fiables ».

Nous indiquons ci-dessous les critères utilisés pour argumenter chacun des points abordés :

Discussions	Critères
Choisir un traitement en fonction du paramètre « prise en compte du nombre de classes »	Stress de Kruskal Tq bien représentés <sup>19</sup> Topologie Groupements fiables
Isoler des groupements fiables	Tq supportant une arête interne Tq mal représentés impliquant un objet
Identifier la structure des groupements fiables	Objets associés Longueur des arêtes externes Verbalisations
Identifier le statut des objets formant un groupement fiable	Objets associés Objet formant une classe unique Sous-classes Longueur des arêtes externes
Identifier des objets spécifiques	Objets associés Objet formant une classe unique Verbalisations
Analyser les objets appartenant à des groupements peu fiables	Objets associés Verbalisations

<sup>19</sup> Tq est mis pour *taux de quadruplets*.

#### 4.1. Tenir compte ou pas du nombre de classes

Pour décider s'il est préférable de tenir compte ou pas du nombre de classes, on utilise d'abord les deux critères suivants : le stress et le taux de quadruplets bien représentés, soit un critère métrique et un critère topologique. Nous laissons de côté les autres critères métriques, souvent redondants par rapport au calcul du stress.

Expérience	Stress	Expérience	Taux de quadruplets bien représentés
<b>anoms-é-00-</b>	<b>0,0713</b>	<b>anoms-é-00-</b>	<b>0,7753</b>
anoms-é-00+	0,0737	anoms-é-00+	0,7610
<b>anoms-p-01-</b>	<b>0,1033</b>	<b>anoms-p-01-</b>	<b>0,7368</b>
anoms-p-01+	0,1041	<b>anoms-p-01+</b>	<b>0,7374</b>
<b>étiq-é-00-</b>	<b>0,0750</b>	<b>étiq-é-00-</b>	<b>0,8327</b>
étiq-é-00+	0,0764	étiq-é-00+	0,8253
<b>étiq-p-01-</b>	<b>0,0722</b>	étiq-p-01-	0,8170
étiq-p-01+	0,0733	<b>étiq-p-01+</b>	<b>0,8407</b>
<b>snoms-é-00-</b>	<b>0,0883</b>	<b>snoms-é-00-</b>	<b>0,7151</b>
snoms-é-00+	0,0911	snoms-é-00+	0,6962
<b>snoms-p-01-</b>	<b>0,0939</b>	<b>snoms-p-01-</b>	<b>0,6871</b>
snoms-p-01+	0,0994	snoms-p-01+	0,6813
<b>snoms-p-98-</b>	<b>0,0714</b>	snoms-p-98-	0,7662
snoms-p-98+	0,0722	<b>snoms-p-98+</b>	<b>0,7841</b>

Tableau 6 : valeur du stress et valeur du taux de quadruplets pour toutes les expériences

En gras, nous avons indiqué quelle était du point de vue du critère examiné la meilleure des deux représentations obtenues.

Les deux traitements ne montrent pas un grand écart. Il est au maximum de 0,0055 pour le stress, de 0,0237 pour le taux de quadruplets bien représentés. Du point de vue du stress, ce sont les traitements « - » qui présentent systématiquement les meilleurs résultats ; du point de vue du taux de quadruplets bien représentés, les résultats sont plus partagés (4 traitements « + » contre 3 traitements « - »).

Les deux critères examinés ne permettent pas de trancher en faveur de l'un ou l'autre traitement. On peut alors examiner les topologies construites et voir si elles présentent des différences notables du point de vue de la structuration des objets. Nous nous limiterons aux « groupements fiables », *i.e.* les groupements isolés sur la base de critères ayant une valeur élevée (cf. § 4.3.). La comparaison portera sur le nombre de groupements fiables et sur le nombre d'objets structurés<sup>20</sup>.

<sup>20</sup> Il faut noter cependant que l'on aurait pu considérer d'autres points :

- la répartition du nombre de classes. On observe en effet dans « anoms-é-00 » que les sujets ont réalisé au moins 3 classes et au plus 11 ; dans « snoms-é-00 » les sujets ont réalisé au moins 4

Seules deux expériences ne présentent pas les mêmes groupements fiables selon que l'on a tenu compte ou pas du nombre de classes :

anoms-é-00-	anoms-é-00+	snoms-é-00-	snoms-é-00+
orange, citron	orange, citron	ail, poisson	ail, poisson, fuel
fuel, poisson, ail	fuel, poisson, ail	orange, citron	
lavande, violette	lavande, violette		
vanille, cannelle	vanille, cannelle		
eucalyptus, menthe			

Tableau 7 : liste des groupements fiables, expériences « anoms-é-00+/- » et « snoms-é-00+/- »

On observe les deux cas de figure suivants :

- ajout/suppression d'un groupement au profit de l'un des deux traitements : « eucalyptus, menthe » dans « anoms-é-00- » n'est pas construit dans « anoms-é-00+ » ; « orange, citron » dans « snoms-é-00- » n'est pas construit dans « snoms-é-00+ ». Il s'avère ici que ce sont les traitements qui n'ont pas tenu compte du nombre de classes qui sont les plus structurants ;
- modification de la structuration d'un groupement : dans « snoms-é-00+ », le fuel est agrégé au groupement « poisson, ail ». C'est le traitement « + » qui a ici l'avantage.

Si on adopte un point de vue quantitatif strict (nombre de groupements fiables et nombre d'objets impliqués), les traitements « - » ont alors l'avantage.

### **Bilan**

Les traitements qui ne tiennent pas compte du nombre de classes semblent légèrement supérieurs. Pour cette raison, mais aussi pour des raisons de simplicité d'exposition, nous nous attacherons par la suite uniquement aux traitements « - ». Nous ne ferons donc plus mention des signes « - » et « + » dans le nom des traitements.

## **4.2. Analyse globale**

Au vu des critères métrique et topologique, ces différentes expériences ne se laissent pas aisément représenter par une distance d'arbres :

- le stress est toujours supérieur à 5 % ; il atteint même plus de 10 % dans certaines expériences ;
- le taux de quadruplets bien représentés n'atteint 80 % que pour l'expérience réalisée sans stimuli olfactifs (« étiquettes »). Les expériences « snoms »

---

classes et au plus 9. En fonction d'hypothèses ou de résultats externes (par exemple, si l'on dispose de résultats sur le nombre attendu de classes), on pourrait alors préférer tel ou tel traitement ;

- les objets impliqués dans les classes faisant problème si l'on dispose d'hypothèses spécifiques ;
- les verbalisations qui y sont associées, etc.

présentent les moins bons résultats. On notera cependant que les résultats de 1998 sont nettement meilleurs que les résultats de 2000 et 2001<sup>21</sup> ;

- on observe également une grande variation quand on compare les résultats des différentes expériences : un éparpillement des groupes et des variations importantes dans les distances entre certains odorants.

Les résultats de l'expérience « snoms-p-98 » ont fait l'objet d'une étude particulière (Guénoche & Garreta 2001), qui montre que (i) la topologie construite correspond à une distance d'arbres bruitée de l'ordre de 20 % ; (ii) tout autre modèle de distances appliqué à ces données fournirait de moins bons résultats qu'un modèle de distance d'arbres.

Il n'en reste pas moins que les critères ont des valeurs relativement faibles. Aussi, par prudence, on s'intéressera non pas à la « structure d'arbre » dans sa totalité, mais aux « groupements », *i.e.* des classes dont nous discutons le statut catégoriel<sup>22</sup>. On sera amené ainsi à distinguer des « groupements fiables », isolés sur la base de critères ayant une valeur élevée ; les autres groupements formeront les « groupements peu fiables » (cf. § 4.6.).

### 4.3. Isoler des groupements fiables

Pour isoler des groupements susceptibles de former des catégories, on se servira du taux de quadruplets supportant chaque arête interne. Le taux de quadruplets mal représentés impliquant un objet servira à une validation globale des résultats.

On considérera comme « groupements fiables » les groupements qui s'opposent aux autres objets de la topologie par une arête interne supportée par un taux de quadruplets supérieur ou égal à 0,90. On appellera « groupements peu fiables » les groupements qui ne remplissent pas cette condition. Le choix de la valeur « plancher », ici 0,90, est arbitraire. C'est une valeur élevée dans l'intervalle [0,1]. Soit l'exemple suivant, tiré de « snoms-p-01 » :

---

<sup>21</sup> Que l'on pourrait mettre au compte des conditions de passation de l'expérience. Cf. note 10.

<sup>22</sup> Nous distinguons ainsi (i) la classe : qui relève d'une définition mathématique ; (ii) la catégorie : qui relève d'une définition psychologique (mais on parlera aussi de catégorie linguistique lors de l'analyse des verbalisations). La notion de groupement vise ainsi l'objet même de notre discussion : une hypothétique catégorie.

**Taux de quadruplets supportant chaque arête interne**

arête 17 -- 19 :9	/	13	=	0,6923
arête 18 -- 20 :10	/	13	=	0,7692
arête 19 -- 28 :60	/	60	=	1,0000
arête 20 -- 28 :58	/	60	=	0,9667
arête 21 -- 25 :10	/	13	=	0,7692
arête 22 -- 24 :8	/	13	=	0,6154
arête 23 -- 26 :7	/	13	=	0,5385
arête 24 -- 29 :62	/	84	=	0,7381
arête 25 -- 27 :34	/	60	=	0,5667
arête 26 -- 27 :20	/	60	=	0,3333
arête 27 -- 29 :76	/	189	=	0,4021
arête 28 -- 30 :23	/	81	=	0,2840
arête 29 -- 30 :43	/	108	=	0,3981

Les arêtes supportées par un taux de quadruplets supérieur ou égal à 0,90 sont les arêtes 19-28 et 20-28.

- Le nœud 19 implique le nœud 17 et le fuel. Le nœud 17 quant à lui rassemble l'ail et le poisson. On considérera donc le groupement « ail, poisson, fuel » comme un groupement fiable. Comme l'arête interne 17--19 est supportée par un taux de quadruplets inférieur à 0,90 (0,6923), on ne considérera pas que ce groupement comporte un sous-groupement fiable.
- Le nœud 20 implique le nœud 18 et l'anis. Le nœud 18 quant à lui rassemble l'orange et le citron. On considérera donc le groupement « orange, citron, anis » comme un groupement fiable. De même que précédemment, comme l'arête interne 18--20 est supportée par un taux de quadruplets inférieur à 0,90 (0,7692), on ne considérera pas que ce groupement comporte un sous-groupement fiable.

Rappelons que la méthode de construction joint à chaque itération **exactement** deux objets. Aussi, dans une situation où trois objets sont très proches, deux d'entre eux seront d'abord rassemblés ; puis à l'itération suivante, le troisième sera agrégé. L'arête interne entre les deux premiers objets et le troisième objet peut donc constituer une « fausse arête ». Ne pas tenir compte de l'ordre de construction présente l'intérêt de pouvoir retenir des groupements dont la configuration interne n'est pas fiable ou moins fiable (cf. taux peu élevés des arêtes 17--19 et 18--20), alors qu'ils constituent des groupements fiables en regard du reste de la topologie.

Si on applique cette méthode aux résultats des différentes expériences, les groupements fiables qui se retrouvent quasiment dans toutes les expériences sont « fuel, poisson, ail » et « orange, citron »<sup>23</sup>.

<sup>23</sup> Dans « snoms-é-00 », on a « ail, poisson » ; dans « snoms-p-01 », on a « orange, citron, anis » mais pas de sous-groupement « orange, citron ».

Une manière de valider ces résultats est d'observer la valeur du taux de quadruplets mal représentés impliquant chacun des odorants des groupements considérés. Nous présentons ci-dessous le taux maximum observé, le taux minimum observé, la moyenne des taux pour chacune des expériences présentant ces groupements fiables, puis le taux associé à chacun des odorants :

	anoms-é-00	anoms-p-01	étiq-é-00	étiq-p-01	snoms-p-98
<b>Maximum</b>	31,6	38,9	25,9	33	40,4
<b>Minimum</b>	14,3	18,9	7,3	11,4	17,4
<b>Moyenne</b>	<b>21,9</b>	<b>25,9</b>	<b>16,6</b>	<b>18</b>	<b>23,3</b>
<b>fuel</b>	18,9	18,9	14,9	18	17,8
<b>poisson</b>	19,1	20,9	12,7	17,6	18,7
<b>ail</b>	21,1	21,3	18,7	13,6	18,5
<b>orange</b>	24,6	20,2	7,9	13,8	19,6
<b>citron</b>	25,5	20,7	7,3	16,5	20,9

Tableau 8 : taux de quadruplets mal représentés

Les odorants retenus présentent des taux proches de la moyenne observée. Aucun ne présente un taux très élevé (toujours inférieur à 26 %). On peut ainsi considérer que les groupements identifiés sont validés. Précisons qu'il s'agit ici d'une validation globale. En effet, on ne peut utiliser le taux de quadruplets mal représentés impliquant un objet comme un indicateur direct de la fiabilité des odorants impliqués dans le groupement, car il n'y a pas de corrélation simple entre un objet impliqué dans un groupement fiable et le taux de quadruplets mal représentés impliquant cet objet : un objet appartenant à un groupement fiable peut présenter un taux élevé ; un objet appartenant à un groupement peu fiable peut présenter un taux faible.

Une fois que l'on a isolé des groupements fiables, il est intéressant d'observer :

- (i) s'il existe des groupements fiables de type différent ;
- (ii) le statut des objets constituant un groupement fiable ;
- (iii) les objets impliqués dans les groupements peu fiables.

Ces différents points font respectivement l'objet des sections suivantes.

#### 4.4. Structuration des groupements fiables

Nous discuterons dans cette section des expériences dans lesquelles on a identifié les deux groupements fiables « fuel, poisson, ail » et « orange, citron », soient les expériences « anoms00 », « anoms-p-01 », « snoms-p-98 », « étiq-é-00 » et « étiq-é-01 ». Pour l'analyse des verbalisations, nous laisserons de côté « étiq-é-00 », expérience pour laquelle nous ne disposons pas des descriptions des classes réalisées.

Pour inférer une différence de structuration, on observera :

- le nombre d'objets associés au groupement ;
- la longueur des arêtes externes ;
- les verbalisations.

##### *Le nombre d'objets associés au groupement*

On considère ici le nombre d'objets associés, *i.e.* les objets n'appartenant pas à la classe produite par Addtree, mais qui ont été associés par les sujets qui ont constitué cette classe (cf. note 17) :

Expérience	Nb d'odorants associés à « fuel, poisson, ail »	Nb d'odorants associés à « orange, citron »
anoms-é-00	0	31
anoms-p-01	5	22
snoms-p-98	1	34
étiq-é-00	0	37
étiq-p-01	0	21

Tableau 9 : nombre d'odorants associés aux groupements fiables

Les odorants fuel, poisson et ail sont clairement traités de manière différente si on les compare à l'orange et au citron. Ils forment une « classe quasi fermée » attirant très peu d'autres odorants.

### ***La longueur des arêtes externes***

On observe les résultats suivants :

Expérience	Longueur des arêtes externes				
	fuel	poisson	ail	orange	citron
anoms-é-00	19,55	17,29	23,74	1,79	3,48
anoms-p-01	13,16	13,16	22,81	5,64	4,89
snoms-p-98	35,86	21,70	20,80	21,92	23,08
étiq-é-00	38,73	32,01	25,13	2,38	2,38
étiq-p-01	20,89	19,11	29,32	9,55	10,45
<b>Moyenne</b>	<b>25,64</b>	<b>20,65</b>	<b>24,36</b>	<b>8,26</b>	<b>8,86</b>

Tableau 10 : longueur des arêtes externes des odorants des groupements fiables

Ce critère permet lui aussi de séparer les deux groupements considérés. Les arêtes externes de l'ail, du poisson et du fuel sont en moyenne beaucoup plus élevées que celles du citron et de l'orange.

### ***Les verbalisations***

On rassemble les verbalisations des sujets qui ont réalisé le groupement produit par Addtree, soient les descriptions des classes comprenant au moins « fuel, poisson, ail », et des classes comprenant au moins « orange, citron ».

### ***Analyse catégorielle des descriptions***

La première analyse porte sur la catégorie des descriptions utilisées. Le plus souvent de catégorie nominale et adjectivale, quelques expressions sont de nature prépositionnelle (*sans odeur*), verbale (*sent mauvais*), phrastique (*ça se mange*) ou sont des relatives (*qui puent*). Nous avons ainsi comptabilisé toutes les

occurrences de ces catégories, un même sujet pouvant avoir employé plusieurs expressions de la même catégorie ou de catégories différentes<sup>24</sup> :

Expériences : anoms-é-00, anoms-p-00, étiq-p-01, snoms-p-98				
Groupement	Exp. nominales (occ. et %)	Exp. adjectivales (occ. et %)	Exp. verbales et phrastiques (occ. et %)	Total (occ. et %)
<b>fuel, poisson, ail</b>	23	29	8	60
<b>Nb sujets : 39 (sur 98)</b>	38,33	48,33	13,33	100
<b>orange, citron</b>	68	55	6	129
<b>Nb sujets : 73 (sur 98)</b>	52,71	42,64	4,65	100

Tableau 11 : répartition catégorielle des verbalisations, I

On voit apparaître une première différence : « fuel, ail, poisson » est construit prioritairement avec des expressions adjectivales ; « orange, citron » est construit avec des expressions nominales et adjectivales.

La différence catégorielle s'accroît dès que l'on ne tient pas compte des expressions comportant un terme générique d'odeur (*odeur, parfum et senteur* (dans leur sens non artefactuel), etc.) et du verbe *sentir*<sup>25</sup>. En effet, si l'on opère un classement référentiel, ces termes sont peu informatifs<sup>26</sup> :

Expériences : anoms-é-00, anoms-p-00, étiq-p-01, snoms-p-98				
Groupement	Exp. nominales (occ. et %)	Exp. adjectivales (occ. et %)	Exp. verbales et phrastiques (occ. et %)	Total (occ. et %)
<b>fuel, poisson, ail</b>	7	56	3	66
<b>Nb sujets : 39 (sur 98)</b>	10,61	84,85	4,55	100
<b>orange, citron</b>	42	92	3	137
<b>Nb sujets : 73 (sur 98)</b>	30,7	67,2	2,19	100

Tableau 12 : répartition catégorielle des verbalisations, II

Le groupement « fuel, ail, poisson » est construit avec environ 8 fois plus d'expressions adjectivales que d'expressions nominales ; pour « orange, citron », le rapport est de l'ordre de 2.

### Analyse des types de termes

<sup>24</sup> Dans l'analyse des verbalisations, les pourcentages sont donnés à titre indicatif, en ce qu'ils favorisent parfois une meilleure lisibilité des résultats. Ils sont arrondis au demi-point le plus proche. *Occ.* est mis pour *occurrence*, *exp.* pour *expression*.

<sup>25</sup> On prend alors en compte le groupe adjectival ou le groupe prépositionnel qui suit. Par ex. *mauvaise* dans *odeur mauvaise* ; *nourriture* dans *odeur de nourriture*.

<sup>26</sup> Le nombre total d'occurrences diffère entre les tableaux 11 et 12. Cela s'explique par le fait que certaines expressions comportent des expansions coordonnées : par ex. *odeur agréable et légère*. Dans le tableau 11, on prend en compte uniquement *odeur* ; dans le tableau 12, on comptabilise les deux adjectifs.



A partir de l'analyse en termes de catégories, on peut procéder à une analyse par types. On a ici classé les expressions nominales, éléments tête de la description<sup>27</sup> :

Expériences : anoms-é-00, anoms-p-00, étiq-p-01, snoms-p-98						
Groupement	Propriété	Nom générique d'odeur/goût	Objet naturel	Objet artefactuel	Événement	Total
fuel, poisson, ail	occ.	2	17	0	2	23
	%	8,70	73,91	0	8,70	100
orange, citron	occ.	4	33	17	1	68
	%	5,88	48,53	25	1,45	100

Tableau 13 : répartition des verbalisations par types

- « fuel, poisson, ail » est plus souvent identifié par un terme générique d'odeur et des propriétés, quasiment jamais par des objets. Nous y reviendrons dans le bilan ci-dessous.

- « orange, citron » est identifié quasi à égalité par des termes génériques et des objets. Les termes dénotant des objets naturels sont un peu plus nombreux que ceux qui dénotent des objets artificiels.

On peut ainsi mettre en évidence deux stratégies d'identification et de regroupements des odeurs :

- d'une part, celle qui consiste à vouloir identifier les odeurs/odorants par une dénomination dénotant un objet naturel, qui conduit à une classification taxinomique des objets naturels (par exemple, catégorie des fruits, des fleurs, des plantes, etc.) ;
- - d'autre part, celle qui repère les odeurs/odorants au travers de pratiques et d'activités<sup>28</sup>, dans lesquels les odeurs interviennent. On conserve dans cette stratégie une identification en termes d'objets, peut-être jugés comme ayant une réalité plus objective, mais aussi comme métonymie de l'activité (*plats, gel douche*).

A cela il faut noter les deux points suivants :

- (i) ces stratégies ne sont pas équivalentes : les sujets jugent la première meilleure que la seconde ; et dans la première, ils jugent l'identification meilleure quand

<sup>27</sup> Les types :

- « propriété » regroupe des termes tels que *amertume, astringence* ;  
 - « objet naturel » : on admet ici l'hypothèse non démontrée de Putnam (1975) que les objets naturels forment un type spécifique ; on y a rangé les termes tels que *orange, citron, fruit, fleur*.  
 - « objet artificiel » correspond au type « objet matériel » de Godard et Jayez (1996) ; on y a rangé les termes tels que *cire pour bois, sels de bain* ;  
 - « événement » regroupe les événements tels qu'ils ont été définis par Godard et Jayez (1996) ; on y a rangé les termes tels que *décomposition*.  
 - on a fait apparaître dans la colonne « nom générique d'odeur/goût » les expressions dénotant explicitement une odeur ou un goût, telles que *odeur mauvaise, odeur piquante*.

<sup>28</sup> L'importance des pratiques et des activités a aussi été montrée dans la catégorisation des couleurs, cf. Dubois et Grinevald (2000).

elle dénote un objet visant le niveau de base d'identification des objets visuels (*citron, orange, etc.*)<sup>29</sup>.

- (ii) ces deux stratégies entretiennent des liens étroits. En effet, si les sujets sont satisfaits d'une identification par des noms dénotant des objets naturels, cette identification ne fournit pas cependant un seul classement, taxinomique, qui supposerait que tous les odorants, ou une grande partie, aient été identifiés de la sorte. Les pratiques et les activités dans lesquelles ces odeurs sont repérées interviennent en concurrence de la taxinomie des objets naturels, les choses se compliquant singulièrement quand un même odorant est considéré comme pouvant relever de différentes pratiques<sup>30</sup>.

### **Analyse référentielle**

On prend en compte ici toutes les expressions quelle que soit leur catégorie. Leur analyse a conduit à distinguer des descriptions associées à des propriétés : le jugement hédonique (JH +/- positif ou négatif), l'intensité (I), la familiarité (F), la typicité (T) ; et des descriptions associées à différents domaines de connaissances : la décomposition (D), la nourriture (N), la chimie (Ch), le végétal (V), les cosmétiques (Co) et les produits d'entretien (P)<sup>31</sup>. On obtient :

---

<sup>29</sup> C'est ce que l'on observe dans les discours des sujets dans l'expérience « snoms-p-98 ». Les odorants identifiés par un nom dénotant un objet naturel relevant du niveau de base de l'identification des objets visuels ne font pas l'objet de commentaires. Par contre, nombreux sont les commentaires dans les autres cas. Par exemple, dans la tâche d'identification des odorants :

« Objets naturels »  
- orange : Ouais, c'est le plus proche des agrumes que possible. C'est ni l'orange, ni le. C'est l'orange amère ou quelque chose comme ça. (sujet 16)  
- orange : Un fruit. Que je connais. C'est un. Ouais, disons, une mandarine. (sujet 37)  
- orange : Allez, je vais dire le 9 agrume, que je sais pas le caractériser. (sujet 24)  
« Objets artefactuels »  
- anis : Moi, je dirais pastis, mais bon. (sujet 12)  
- citron : Ben, ça, c'est. Un parfum, je veux dire, un peu léger le parfum, mais je saurais pas dire plus précisément. (sujet 38)  
- violette : Je, on dirait une odeur un peu lourde d'un parfum, d'une eau de toilette pas bon marché ; [...] le mimosa. C'est pas, il est un peu lourd, c'est un peu lourd comme, comme parfum. Mais, c'est une odeur que j'appelle mimosa. (sujet 11)

<sup>30</sup> Ainsi dans les discours des sujets ayant passé l'expérience « snoms-p-98 » :

- sujet 4, à propos de l'identification de la lavande : [...] ça sent la *fleur*, mais en même temps, une *fleur*, je ne sais pas laquelle, vers le *savon*.  
- sujet 11 : Le groupe 1, 6, c'est plutôt des *odeurs de parfums*, c'est-à-dire de *fleurs*. Pas vraiment, puisque là il s'agit de vanille, mais bon, c'est. 1, c'est franchement, j'aurais pu le mettre dans ce groupe-là, mais 6, c'est franchement une *odeur un peu lourde de fleur*. Mais maintenant, c'est vrai que la vanille intervient dans les *eaux de toilette*. Alors, c'est peut-être parce que j'ai une eau de toilette à la vanille que je l'ai mise là.  
- sujet 35, à propos de l'identification de la violette : C'est vraiment un *savon*. Je dirais, c'est la *rose*. Il y a des *bonbons* comme ça aussi. Oui une *fleur*.

<sup>31</sup> - « Jugement hédonique » (JH) : rassemble l'ensemble des expressions dénotant la propriété de plaisir (+) ou de déplaisir (-). Par exemple, *bon, mauvais, écœurant, dégoût*.

- « Intensité » (I) : rassemble l'ensemble des expressions dénotant la propriété d'intensité. Par exemple, *léger, intense, fort*.

Expériences : anoms-é-00, anoms-p-00, étiq-p-01, snoms-p-98										
Groupement	JH +	JH -	I	F	D	N	Ch	Autres	Total	
fuel, poisson, ail occ.	1	50	4	1	4	2	1	3	66	
%	1,52	75,76	6,06	1,52	6,06	3,03	1,52	4,55	100	

Expériences : anoms-é-00, anoms-p-00, étiq-p-01, snoms-p-98												
Groupement	JH +	JH -	I	F	T	V	N	Co	P	Ch	Autres	Total
orange, citron occ.	21	6	9	2	1	29	46	7	3	4	9	137
%	15,33	4,38	6,57	1,46	0,73	21,17	33,58	5,11	2,19	2,92	6,57	100

Tableau 14 : répartition référentielle des verbalisations

La répartition référentielle vient confirmer la différence de type entre les deux groupements :

- - « fuel, poisson, ail » constitue le groupement des « odeurs désagréables » : il est visé très majoritairement par des adjectifs dénotant un jugement négatif<sup>32</sup>. Les sujets sont aussi d'accord pour les juger fortes et/ou les associer aux déchets, à la décomposition.
- - « orange, citron » est un groupement d'odeurs perçues comme agréables<sup>33</sup> et de faible intensité. Si une majorité se dégage autour du thème de la nourriture, auquel on peut ajouter les végétaux (spécifiés majoritairement comme *fruit*,

---

- « Familiarité » (F) : rassemble l'ensemble des expressions dénotant le caractère familier de l'odorant. Par exemple, *familier, peu coutumier*.

- « Typicité » (T) : rassemble l'ensemble des expressions dénotant le caractère (proto)typique de l'odorant. Par exemple, *caractéristique (du nom)*. Ces expressions se trouvent essentiellement dans les expériences « anoms ».

- « Décomposition » (D) : rassemble l'ensemble des expressions dénotant des déchets, une décomposition organique. Par exemple, *poubelle, avarié*.

- « Nourriture » (N) : rassemble l'ensemble des expressions dénotant des objets comestibles et les activités/lieux qui y sont rattachées et/ou des propriétés renvoyant au goût. Par exemple, *cuisine, plats, sorbet d'été, bonbon, fruit, ça se mange, culinaire, salé*.

- « Chimie » (Ch) : rassemble l'ensemble des expressions dénotant une origine chimique. Par exemple, *chimie, chimique, artificielle, synthétique*.

- « Végétal » (V) : rassemble l'ensemble des expressions dénotant des plantes, quel que soit le degré de spécification. Par exemple, *végétal, plante, fleur, fruit, pomme, orange*.

- « Cosmétique » (Co) : rassemble l'ensemble des expressions dénotant des objets associés aux cosmétiques, aux soins corporels. Par exemple, *parfum, sels de bain, savon*.

- « Produit d'entretien » (P) : rassemble l'ensemble des expressions dénotant des objets utilisés dans les tâches ménagères. Par exemple, *cire pour bois, cif*.

- « Autres » : rassemble l'ensemble des expressions n'entrant pas dans une des catégories précédentes. Il s'agit essentiellement, pour les expériences à consigne émotionnelle, des expressions dénotant des évocations variées (*soleil, souvenirs d'enfance*). Elles apparaissent de manière marginale dans les groupements étudiés.

<sup>32</sup> Le seul sujet qui a utilisé un adjectif positif (*suave*) décrit aussi son groupement comme *écaeurant, irritant*.

<sup>33</sup> Un seul sujet note explicitement que le groupement est *désagréable*. Les 5 autres sujets l'ont noté *piquant* ou *un peu piquant*.

*agrume, orange, citron, etc.*), on remarque d'autres domaines plus faiblement représentés, tels que les cosmétiques, la chimie et les produits d'entretien.

### **Bilan**

Les trois critères que nous nous sommes donnés (nombre d'objets associés, longueur des arêtes externes, analyse des verbalisations) sont intéressants en ce qu'ils permettent d'inférer qu'on a affaire à des groupements de type différent :

- « fuel, poisson, ail » constitue ainsi une « classe quasi fermée » ; il est peu associé à d'autres odorants ; les odeurs/odorants qui le composent ne se ressemblent que sur la base d'une propriété, le caractère hédonique, dans son versant négatif. On peut alors émettre l'hypothèse que les odeurs désagréables, à la différence des odeurs agréables, n'ont pas besoin d'être identifiées, du fait précisément de leur caractère désagréable<sup>34</sup>. On peut alors expliquer la quasi-absence de noms dénotant des objets (cf. tableau 13). Il apparaît du même coup que (i) le caractère hédonique négatif est un jugement de niveau de base et (ii) que le caractère hédonique ne constitue par un jugement dimensionnel (ce qui a été confirmé d'un autre point de vue dans les travaux de Rouby & Bensafi 2002 ; Bensafi *et al.* 2002a, b, c) ;

- « orange, citron » constitue quant à lui une « classe ouverte » sur les autres odorants. Etant donné la longueur des arêtes externes, on peut en inférer que les deux odorants présentent des caractéristiques olfactives similaires, ce qui peut être aussi confirmé à partir de l'analyse des verbalisations. S'il y a consensus sur le jugement d'intensité, il est un peu moins affirmé pour le jugement hédonique. Quant aux domaines de connaissances associés, ils sont plus diversifiés, même si celui de la nourriture apparaît majoritairement. Nous reviendrons sur ce point lors de l'analyse des verbalisations des « groupements peu fiables » (cf. § 4.6.).

### **4.5. Le statut des objets constituant un groupement**

Dans cette section, on examine si les objets constituant un groupement ont des statuts différents. A partir de cette analyse, on peut poser des hypothèses sur la structuration catégorielle des groupements, sur l'identification d'un prototype, sur la constitution du groupement selon une échelle de typicité. Mais, étant donné la complexité de ces questions d'une part<sup>35</sup>, et les données et les traitements qui

---

<sup>34</sup> C'est explicitement mentionné par le sujet 25 dans « snoms-p-98 », à propos de l'identification de la cannelle : « C'est pas agréable, ça pue ça ; c'est pas agréable. Je donne pas de nom, ça m'énerve. »

<sup>35</sup> Ces questions ont été travaillées et d'un point de vue psychologique (cf. Rosch 1976, 1978 ; Dubois 1991) et d'un point de vue linguistique (Lakoff 1987 ; Kleiber 1990 ; Dubois 1991), dans des perspectives qui n'ont pas toujours permis de clarifier le débat. Elles engagent en outre des positions philosophiques, notamment sur la question du « mode d'existence » des catégories et du prototype, selon que l'on considère ou pas qu'il s'agit de représentations conceptuelles soumises à des ordres différents de détermination (Dubois 1991). Enfin, l'importance même de certaines questions a évolué. Ainsi, si l'identification du prototype dans les premières théorisations de Rosch (Rosch *et al.* 1976) était une question centrale, elle est devenue moins fondamentale quand on a proposé des structurations catégorielles organisées selon une ressemblance de famille (Rosch

servent d'illustrations d'autre part<sup>36</sup>, nous nous attacherons à une seule question, celle de la structuration catégorielle.

Nous allons examiner à nouveau le groupement « fuel, poisson, ail »<sup>37</sup> à partir des critères suivants :

- les objets formant une classe unique ;
- les objets associés ;
- les classes constituées par les sujets qui n'ont pas constitué le groupement considéré (qui forment alors une sous-classe du groupement considéré) ;
- la longueur des arêtes externes.

***Les objets formant une classe unique***

---

1978 ; Lakoff 1987 ; Kleiber 1990 ; Dubois 1991), le prototype n'étant alors plus réputé exister pour toutes les catégories. Cependant, on identifie parfois un élément « premier », dont on ne sait (les critères d'identification restant flous) s'il est premier dans l'ordre de la découverte ou s'il est premier parce qu'il joue un rôle particulier.

<sup>36</sup> Dans le type de représentation produit par Addtree, Barthélemy (1991) a fait l'hypothèse que le nœud rassemblant l'ensemble des éléments de la classe pouvait constituer le prototype. Mais l'algorithme qu'il avait retenu diffère de celui que nous avons utilisé : la construction de groupements comprenant plus de deux objets y était possible. Ce point est déterminant. Dans un groupement comprenant plus de deux objets, l'identification du prototype est en effet plus complexe :

- parce que plusieurs nœuds sont possibles ;  
- parce que le calcul de la distance d'un objet au prototype doit prendre en compte des longueurs d'arêtes internes, éventuellement supportées par des taux de quadruplets faibles sans être nuls, etc.  
On peut considérer que le prototype partage :

(i) soit plus de propriétés avec le dernier objet agrégé, ce qui n'est pas forcément un résultat souhaitable (étant donné la similarité des longueurs des arêtes externes, la distance du dernier odorant agrégé au prototype supposé se trouve, dans les expériences rapportées ici, quasiment toujours être inférieure à la distance d'un odorant agrégé auparavant, car cette distance met en cause et l'arête externe et l(es) arête(s) interne(s)) ;

(ii) soit plus de propriétés avec l'un des premiers objets agrégés (mais là encore, dans les données présentées ici, étant donné la similarité des longueurs des arêtes externes, aucun odorant n'est distinguable) ;

(iii) soit plus de propriétés avec un objet agrégé ni en premier ni en dernier, et dans ce cas, on manque de critères.

Enfin, travailler cette question en se servant des indications fournies par les sujets ne résout pas le problème. L'objet le plus typique qui aura été mentionné peut ne pas faire partie des objets constituant le groupement fiable. Ces indications ne sont pas alors de ce point de vue directement utiles.

<sup>37</sup> On utilisera les résultats des expériences considérées précédemment et ceux de « snoms-p-01 » (qui n'avaient pas été pris en compte dans les comparaisons menées dans la section § 4.4., du fait de l'absence du groupement fiable « orange, citron »).

On observe pour les trois odorants les résultats suivants :

Expérience	Nb sujets	Nb fois où l'odorant forme une classe unique		
		fuel	poisson	ail
anoms-é-00	19	6	4	3
anoms-p-01	19	3	1	3
snoms-p-98	40	17	6	8
snoms-p-01	24	3	2	2
étiq-é-00	21	11	10	6
étiq-p-01	20	8	5	5
<b>Total</b>	<b>143</b>	<b>48</b>	<b>28</b>	<b>27</b>

Tableau 15 : formation de classe unique (fuel, poisson, ail)

Globalement, le poisson et l'ail restent similaires, le fuel est l'odorant qui se trouve à former le plus souvent une classe unique. Cela reste vrai même si l'on ne tient pas compte des résultats de « snoms-p-98 »<sup>38</sup>.

### **Les objets associés**

On considère ici le nombre d'objets associés à chacun des odorants considérés, autres que ceux avec lesquels il forme le groupement fiable. On obtient les résultats suivants :

Expérience	Nb sujets	Nb d'odorants associés à fuel, poisson et ail (autres que fuel, poisson, ail)		
		fuel	poisson	ail
anoms-é-00	19	3	7	19
anoms-p-01	19	15	11	17
snoms-p-98	40	15	21	26
snoms-p-01	24	37	31	35
étiq-é-00	21	5	5	27
étiq-p-01	20	3	5	26
<b>Total</b>	<b>143</b>	<b>78</b>	<b>80</b>	<b>150</b>

Tableau 16 : nombre d'odorants associés à fuel, poisson et ail (autres que fuel, poisson, ail)

Globalement les résultats pour le fuel et le poisson restent similaires. Seul l'ail manifeste un écart significatif. Cela reste vrai même si l'on ne tient pas compte des résultats des expériences qui montrent le plus grand écart entre l'ail et les deux autres odorants.

### **Les classes constituées ne comprenant pas la totalité du groupement**

<sup>38</sup> Cf. note 10 sur les conditions de passation de « snoms-p-98 ». Seul l'ordre de grandeur de la différence entre le fuel et les deux autres change : 60 % si on tient compte des résultats de « snoms-p-98 » contre 30 % si on n'en tient pas compte.

Certains sujets n'ont pas rassemblé les trois odorants mais seulement deux d'entre eux :

Expérience	Nb sujets	Nb. de sujets ayant réalisé uniquement les classes		
		fuel, poisson	poisson, ail	fuel, ail
anoms-é-00	19	4	2	0
anoms-p-01	19	6	4	0
snoms-p-98	40	4	15	4
snoms-p-01	24	4	5	3
étiq-é-00	21	2	5	1
étiq-p-01	20	7	3	0
<b>Total</b>	<b>143</b>	<b>27</b>	<b>34</b>	<b>8</b>

Tableau 17 : répartition de fuel, poisson, ail, pris 2 à 2

Les groupements qui arrivent en tête impliquent toujours le poisson<sup>39</sup>.

### *La longueur des arêtes externes*

La longueur des arêtes externes est la suivante :

Expérience	Longueur des arêtes externes		
	fuel	poisson	ail
anoms-é-00	19,55	<b>17,29</b>	23,74
anoms-p-01	<b>13,16</b>	<b>13,16</b>	22,81
snoms-p-98	35,86	21,70	<b>20,80</b>
snoms-p-01	19,36	17,11	<b>16,22</b>
étiq-é-00	38,73	32,01	<b>25,13</b>
étiq-p-01	20,89	<b>19,11</b>	29,32
<b>Moyenne</b>	<b>24,59</b>	<b>20,06</b>	<b>23</b>

Tableau 18 : longueur des arêtes externes de fuel, poisson, ail

En gras, nous avons indiqué quel était l'odorant qui montrait la longueur la plus petite. Les trois odorants présentent en moyenne des ordres de grandeur similaires<sup>40</sup>, qui sont toujours largement supérieurs à ceux que l'on a observés pour « orange, citron ». L'odorant présentant la longueur la plus petite est, selon les expériences considérées, l'un des trois odorants (cf. indications en gras).

### *Bilan*

Selon le 1<sup>er</sup> critère, le fuel forme le plus souvent une classe unique. La différence entre le fuel et les deux autres est donc la plus grande. Selon le 2<sup>ème</sup> critère, l'ail est le plus associé avec les autres odorants de l'échantillon. Ce n'est donc pas celui qui est central dans la catégorie des « odeurs désagréables ». Selon le 3<sup>ème</sup> critère, le poisson est le seul à être groupé de manière significative aussi bien avec le fuel que l'ail. On pourrait supposer en première hypothèse que le prototype de cette catégorie présente le plus de propriétés communes avec le poisson (pour les

<sup>39</sup> On notera que c'est l'expérience « snoms-p-98 » qui contribue le plus fortement à la réalisation du groupement « poisson, ail », ce pour quoi nous n'avons pas d'explication.

<sup>40</sup> Rappelons que les distances sont calculées par rapport à 100 : la différence des moyennes est au plus de 4,53.

odorants que nous avons testés). Mais le 4<sup>ème</sup> critère ne distingue aucun de ces odorants : ils sont tous aussi différents les uns des autres et celui qui manifeste la longueur d'arête la plus petite n'est pas toujours le même. On peut ajouter à cela (i) que ces odorants sont rassemblés sur la base d'une seule propriété, le jugement désagréable (cf. § 4.4.) ; (ii) que l'on a constitution d'un sous-groupement fiable, soit « poisson, ail » (« snoms-p-98 »), soit « fuel, poisson » (« anoms-é-00 », « anoms-p-01 », « étiqu-p-01 ») : le sous-groupement fiable observé n'est pas stable.

Ainsi quand on examine le statut des différents éléments constituant ce groupement, aucun ne semble jouer un rôle central et organisateur. Dans ce cas de figure, il semble plus adéquat de poser l'hypothèse d'une organisation catégorielle fondée sur la ressemblance de famille. C'est ce que l'on retiendra pour les « odeurs désagréables ». Au-delà de la nécessaire validation de cette hypothèse par d'autres expériences et d'autres analyses, on voit là la valeur heuristique de la démarche mise en place, en tout cas, pour ce type d'organisation catégorielle. Les critères utilisés ne permettent pas en effet de dire quoi que ce soit à propos de « orange citron »<sup>41</sup>. Comme les résultats des expériences ne fournissent pas un large éventail de tests, il reste, d'un point de vue méthodologique, à éclaircir la véritable portée de ces critères, en les utilisant sur d'autres données. Mais, on sait déjà que cette portée sera limitée, car elle ne couvrira pas « orange, citron », pour lesquels le problème de l'identification de la structure catégorielle et d'un éventuel prototype reste entier.

#### **4.6. Analyse des objets n'appartenant pas à des groupements fiables**

Que peut-on dire des objets ne faisant pas partie de groupements fiables ? On doit distinguer deux cas :

- (i) un objet n'appartient pas à un groupement fiable parce qu'il ne ressemble à aucun autre. Il sera agrégé tardivement dans la construction de l'arbre. Il présentera un nombre peu élevé d'associations avec les autres objets, voire aucune ;
- (ii) un objet n'appartient pas à un groupement fiable parce qu'il ressemble à beaucoup d'autres objets, sans pour autant qu'une configuration n'émerge ; on a tendanciellement un nombre d'associations relativement homogène de cet objet avec les autres.

Nous examinerons successivement ces deux cas, en considérant les odorants suivants : vanille, lavande, eucalyptus, violette, herbe, pomme, cannelle, anis, mûre, chewing-gum, menthe. Ils ne forment jamais ou rarement un groupement fiable, dans les expériences 2000 et 2001.

---

<sup>41</sup> Ce qui est somme toute normal si l'on se rappelle que ces deux odorants sont le plus souvent rassemblés dans une même classe (cf. arêtes internes de faible longueur comparativement aux autres arêtes) et que ce groupement forme une classe ouverte sur les autres odorants (beaucoup d'odorants associés et pas de classes uniques). Quant au critère de la sous-classe, il est inapplicable.



#### 4.6.1. Identification des objets spécifiques

Peut-on retrouver des objets qui auraient des caractéristiques particulières reconnues par tous les sujets et qui, de ce fait, ne ressembleraient à aucun autre ? Si tel est le cas, ils devraient former le plus souvent une classe unique.

On n'observe pas ici un tel type d'objet. Cependant, selon les expériences, certains odorants ont un comportement particulier pour un sous-ensemble des sujets (au moins 15 % des sujets considérés) :

- le chewing-gum : dans les expériences « étiquettes » ;
- - la mûre : dans les expériences à consigne émotionnelle avec stimulus olfactif<sup>42</sup> ;
- la vanille et la cannelle : dans les expériences « anoms ».

Pour le chewing-gum, on peut faire l'hypothèse que c'est le nom choisi qui joue un rôle central. Il dénote un artefact (la quasi-totalité des autres noms dénotent des « objets naturels ») ; l'ingrédient n'y est pas spécifié. Le nom devient problématique dans les expériences « étiquettes », qui ne comportent pas de stimuli olfactifs. A l'appui, on peut signaler qu'il forme une classe unique plus souvent dans les expériences « anoms » que dans les expériences « snoms ».

Pour la mûre, on peut faire l'hypothèse qu'il s'agit d'un problème de détection (connu par ailleurs, cf. Gilbert & Kemp 1996), en particulier dans l'expérience « snoms-é-00 » ; certains disent ne rien sentir, et, dans les verbalisations, on trouve des descriptions telles que *inodore*. Toutefois, cette hypothèse ne semble pas suffisante, dans la mesure où on n'observe pas ce phénomène dans « snoms-p-01 ».

L'explication serait alors à chercher non plus du côté de l'odorant mais du côté de la consigne émotionnelle. Quand on rapporte le nombre de classes uniques au nombre de sujets ayant produit au moins une fois une classe unique, on observe que les sujets face à une consigne émotionnelle produisent en moyenne plus de classes uniques que les sujets face à une consigne perceptive<sup>43</sup>. Le nombre plus élevé de classes ne comportant que la mûre dans « snoms-é-00 » serait alors à mettre au compte de la consigne émotionnelle.

Pour la vanille et la cannelle, ce sont les expériences « anoms » qui sont spécifiques. Nous n'avons pas d'explication, hormis l'hypothèse que ces deux odorants ne constituent pas pour les sujets des odeurs typiques des noms qui ont été choisis.

#### 4.6.2. Analyse des objets appartenant à des groupements peu fiables

---

<sup>42</sup> Mais ce point se vérifie aussi dans l'expérience « snoms-p-98 ».

<sup>43</sup> On peut avancer les deux raisons suivantes : d'une part, les noms proposés ne dénotent pas des émotions. D'autre part, l'étude des verbalisations ne montre pas que les émotions se composent dans une taxinomie partagée. Les étiquettes proposées renvoient principalement à des objets, qui se composent plus aisément dans une taxinomie d'objets naturels (les plantes, les fleurs, les fruits, etc.).

Que le traitement réalisé par Addtree puisse produire des groupements peu fiables ne doit pas étonner. Dans le cadre des expériences menées,

(i) on teste un nombre restreint de 16 odeurs/odorants (qui ne couvrent peut-être pas l'espace catégoriel de manière adéquate) ;

(ii) on n'a pas mesuré précisément les éventuels problèmes de représentativité de l'odorant ;

(iii) quant à la question de la familiarité, si elle se pose, elle n'est pas corrélable<sup>44</sup> de manière simple.

Pour ces trois raisons, la formation de classes conjoncturelles est inévitable. Aussi, du point de vue de l'identification de catégories, les groupements peu fiables ne présentent pas d'intérêt. Par contre, si l'on cherche à expliquer pourquoi tel ou tel objet n'est jamais retenu, pourquoi la catégorisation est fluctuante d'une expérience à l'autre, alors ce type de groupements (re)devient tout à fait intéressant à étudier.

Nous nous limiterons ici aux groupements peu fiables rassemblant des odeurs/odorants reliés par une distance n'impliquant que des arêtes externes ; autrement dit des paires d'objets agrégés, et non des objets agrégés à d'autres groupements :

<b>anoms-p-01</b>	<b>snoms-é-00</b>	<b>snoms-p-01</b>	<b>étiq-é-00</b>
<i>lavande, violette</i> 0,7692	pomme, chewing-gum 0,6923	violette, herbe 0,7692	<b>chewing-gum, menthe</b> 0,6531
<b>vanille, cannelle</b> 0,6154	<i>eucalyptus, menthe</i> 0,5556	<b>chewing-gum, menthe</b> 0,6154	eucalyptus, herbe 0,6250
pomme, mûre 0,8333	<b>vanille, cannelle</b> 0,5758	vanille, pomme 0,5385	
	violette, mûre 0,1667		
	lavande, herbe 0,6364		

Tableau 19 : liste des groupements peu fiables

Sous chaque groupement est indiqué le taux de quadruplets supportant l'arête interne reliant ce groupement au reste de l'arbre.

En gras sont indiqués les groupements identiques dans des expériences différentes ; en italiques, les groupements formant des groupements fiables dans une des autres expériences. On trouve en effet les groupements fiables suivants :

<sup>44</sup> Dans Godinot (1999), on observe que 15 des 16 odorants testés ont reçu une note supérieure 5 (sur une échelle de 10). Si le manque de familiarité peut expliquer le comportement de la mûre (16<sup>ème</sup> position et seule note inférieure à 5), il n'explique pas que d'autres odorants jugés parmi les moins familiers constituent des groupements fiables : par exemple, le citron, le poisson, le fuel, qui arrivent respectivement en 15<sup>ème</sup>, 14<sup>ème</sup> et 13<sup>ème</sup> position. A l'inverse, les odorants jugés parmi les plus familiers ne constituent jamais ou rarement un groupement fiable : par exemple le chewing-gum, l'herbe, la vanille, l'anis, qui arrivent respectivement en 1<sup>ère</sup>, 2<sup>ème</sup>, 3<sup>ème</sup> et 4<sup>ème</sup> position.

- - « vanille, cannelle » dans « anoms-é-00 », « étiqu-é-00 », « étiqu-p-01 » ;
- - « lavande, violette » dans « anoms-é-00 », « étiqu-é-00 » ; comme sous-groupement fiable dans « étiqu-p-01 » ;
- - « eucalyptus, menthe » dans « anoms-é-00 », « anoms-p-01 ».

On privilégiera désormais l'analyse des groupements peu fiables qui, dans au moins une des autres expériences, peuvent apparaître comme groupements fiables, soient les groupements « vanille, cannelle », « lavande, herbe », « eucalyptus, menthe ».

### *Analyse des objets associés*

Si on compare le nombre d'objets associés (et leur nombre d'occurrences) aux odorants impliqués dans des groupements fiables et dans des groupements peu fiables, on ne voit pas apparaître de différences significatives. Les odorants conduisant à un groupement fiable peuvent être même associés à un nombre d'odorants plus élevés que ces mêmes odorants quand ils forment un groupement peu fiable. Par exemple, dans « anoms-é-00 », le groupement fiable « eucalyptus, menthe » est associé 9 autres odorants, pour un total de 20 occurrences, alors que ce même groupement, peu fiable dans « snoms-é-00 » est associé à 6 autres odorants pour un total de 9 occurrences.

Il semblerait en outre qu'on doive rapporter cette comparaison :

- (i) à la nature des odorants : les groupements peu fiables « vanille, cannelle » et « lavande, herbe », dans « anoms-p-01 » sont respectivement associés à 11 et 7 autres odorants ; le premier groupement constitue un groupement plus ouvert que le second ;
- (ii) aux variables de la consigne : dans « anoms-p-01 » et « étiqu-é-00 », soient les expériences avec des noms, les groupements fiables sont plus fortement associés à d'autres odorants que les groupements peu fiables. On observe l'inverse dans « snoms-é-00 » et « snoms-p-01 »<sup>45</sup>.

De ce critère, en tout cas dans le cadre des expériences rapportées ici, on ne peut tirer des oppositions claires entre les groupements fiables et les groupements peu fiables.

### *Analyse des verbalisations*

Nous nous intéresserons ici principalement au groupement « vanille, cannelle » et nous nous limiterons à une analyse catégorielle et une analyse référentielle.

Rappelons que « vanille, cannelle » constitue un groupement peu fiable dans les expériences « anoms-p-01 » et « snoms-é-00 » et un groupement fiable dans l'expérience « anoms-é-00 ». On ne peut donc pas mettre simplement sur le compte des variables des expériences la formation du groupement, puisque ni la consigne émotionnelle/perceptive ni la mise à disposition ou non de noms ne

---

<sup>45</sup> Cette comparaison n'a pu être menée dans le cas des expériences « anoms-é-00 » et « étiqu-p-01 », où aucun groupement peu fiable, constitué d'une seule paire d'objets, n'a été produit.

conduisent systématiquement à la formation de ce groupement. Ce cas permet d'étudier de manière transversale aux différentes expériences le jeu complexe entre deux modes de constitution des groupements, l'un faisant appel au jugement hédonique, l'autre à des domaines de connaissances diversifiés.

D'un point de vue catégoriel, on observe les résultats suivants :

Expérience	Odorants	Nb sujets	Exp. nominales	Exp. adjectivales	Exp. verbales et phrastiques
<b>anoms-p-01</b>	vanille et cannelle	8	7	5	0
<b>anoms-p-01</b>	vanille sans cannelle	11	7	14	1
<b>anoms-p-01</b>	cannelle sans vanille	11	12	11	0
<b>snoms-é-00</b>	vanille et cannelle	8	18	3	0
<b>snoms-é-00</b>	vanille sans cannelle	9	17	1	0
<b>snoms-é-00</b>	cannelle sans vanille	9	15	2	0

Tableau 20 : répartition catégorielle, I

Hormis la différence nette dans la catégorie des expressions (nombre supérieur d'expressions adjectivales pour l'expérience « anoms-p-01 », d'expressions nominales pour l'expérience « snoms-p-00 »), que les sujets aient groupé ou pas les deux odorants considérés, les rapports entre catégories nominales et catégories adjectivales restent similaires, excepté dans « anoms-p-01 » pour la vanille sans la cannelle, qui est visée avec le double d'expressions adjectivales.

Quand on ne tient pas compte des termes génériques des expressions comportant un terme générique d'odeur, on obtient les résultats suivants :

Expérience	Odorants	Nb sujets	Exp. nominales	Exp. adjectivales	Exp. verbales et phrastiques
<b>anoms-p-01</b>	vanille et cannelle	8	2	14	0
<b>anoms-p-01</b>	vanille sans cannelle	11	6	17	1
<b>anoms-p-01</b>	cannelle sans vanille	11	9	16	0
<b>snoms-é-00</b>	vanille et cannelle	8	14	7	0
<b>snoms-é-00</b>	vanille sans cannelle	9	17	1	0
<b>snoms-é-00</b>	cannelle sans vanille	9	15	2	0

Tableau 21 : répartition catégorielle, II

Si on compare les résultats des tableaux 20 et 21, la proportion de descriptions utilisant des termes génériques d'odeurs est plus forte pour le groupement « vanille, cannelle » et pour la cannelle sans la vanille de l'expérience « anoms-p-01 »<sup>46</sup>.

<sup>46</sup> De manière générale, on observe que, face à une consigne émotionnelle, la mise à disposition de noms n'influe pas sur le nombre de descriptions qui utilisent un terme générique d'odeur. En cela les sujets ont globalement respecté la consigne. Par contre, face à une consigne perceptive avec stimulus, on observe que les groupements réalisés sont plus souvent décrits avec un terme générique d'odeur, quand les sujets ont disposé des noms. L'expérience avec consigne perceptive et sans stimulus se situe quant à elle entre les expériences à consigne émotionnelle et les expériences avec stimulus.

Pour la répartition référentielle, on a considéré le jugement hédonique (JH +/-, positif et négatif), le jugement d'intensité (I), la typicité (T), le végétal (V), la nourriture (N), les cosmétiques (Co), les médicaments (M)<sup>47</sup>, la chimie et les produits d'entretien (P). On obtient les résultats suivants :

anoms-p-01		JH +	JH -	I	T	V	N	Co	M	Ch	P	Autres	Total
vanille	occ.	5	2	0	1	1	3	0	0	0	0	4	16
et cannelle	%	31,25	12,5	0	6,25	6,25	18,75	0	0	0	0	25	100
vanille	occ.	6	3	1	0	2	7	1	0	1	0	3	24
sans cannelle	%	26,09	13,04	4,17	0	8,33	29,17	4,17	0	4,17	0	12,5	100
cannelle	occ.	2	5	2	0	7	5	0	0	0	1	3	25
sans vanille	%	8	20	8	0	28	20	0	0	0	4	12	100
snoms-é-00		JH +	JH -	I	T	V	N	Co	M	Ch	P	Autres	Total
vanille	occ.	10	0	0	0	1	8	0	0	0	0	2	21
et cannelle	%	47,62	0	0	0	4,76	38,1	0	0	0	0	9,52	100
vanille	occ.	5	2	0	0	0	6	2	2	0	0	1	18
sans cannelle	%	27,78	11,11	0	0	0	33,33	11,11	11,11	0	0	5,56	100
cannelle	occ.	2	6	0	0	2	2	0	4	0	0	1	17
sans vanille	%	11,76	35,29	0	0	11,76	11,76	0	23,53	0	0	5,88	100

Tableau 22 : répartition référentielle

### ***Le jugement hédonique***

Concernant le jugement hédonique, on observe que :

(i) il constitue la propriété dénotée le plus fréquemment dans chaque groupement observé ;

(ii) les deux odorants suscitent des avis partagés. Ensemble, ils forment toujours un groupement agréable dans « snoms-é-00 », soit un groupement agréable soit un groupement désagréable dans « anoms-p-01 » ; le groupement est alors constitué pour des raisons opposées ;

(iii) le groupement impliquant la cannelle sans la vanille suscite un jugement plus désagréable qu'agréable ; c'est l'inverse pour le groupement impliquant la vanille sans la cannelle ;

dans les deux expériences, la somme de jugements positifs est toujours supérieure à la somme de jugements négatifs.

Etant donné l'importance du jugement hédonique dans le domaine olfactif, on comprend qu'un même sujet trouvant un odorant agréable et l'autre désagréable ne les regroupe pas ensemble. Les deux odorants restent alors séparés sur la base d'une valeur différenciée de ce jugement. C'est par exemple le cas du sujet 5 dans « anoms-p-01 » qui décrit le groupement impliquant la vanille avec *agréable*, *fruité* et celui impliquant la cannelle avec *désagréable*, *piquant* ; le cas du sujet 6 dans « snoms-é-00 » qui décrit le groupement impliquant la vanille avec *plaisir*, *ivresse*, *sensualité* et celui impliquant la cannelle avec *dégoût*, *écœurant*, *répulsion*.

<sup>47</sup> On a regroupé sous cette étiquette les expressions telles que *médicament*, *hôpital*, *maladie*, *pharmacie*.

Le fait que deux odorants suscitent des jugements hédoniques différenciés constitue alors un premier facteur pouvant expliquer le caractère peu fiable du groupement. A l'appui, on peut rappeler que les groupements fiables « fuel, poisson, ail » et « orange, citron » montraient chacun qu'un consensus s'était dégagé autour de ce jugement (respectivement négatif et positif).

Mais la valeur différenciée du jugement hédonique porté sur les odorants considérés n'est pas à elle seule suffisante pour rendre compte des différents cas observés. En effet :

- (i) pour deux odorants donnés, les deux suscitant plus de jugements positifs que de jugements négatifs (ou plus de jugements négatifs que de jugements positifs), on a aussi bien formation de groupements fiables que de groupements peu fiables<sup>48</sup> ;
- (ii) on observe la formation de groupements fiables qui sont décrits à l'aide d'expressions dénotant des jugements positifs *et* des jugements négatifs<sup>49</sup> ;
- (iii) des sujets qui n'auraient pas regroupé deux odorants ne les opposent pas systématiquement selon les deux valeurs de ce jugement<sup>50</sup>.

### ***La multiplicité des domaines de connaissances***

Le deuxième point important est de souligner que les odorants vanille et cannelle sont associés à différents domaines d'activités et de pratiques (cf. tableau 22), qui sont aussi représentés dans les autres groupements peu fiables de chacune des deux expériences :

---

<sup>48</sup> Dans « snoms-é-00 », « vanille, cannelle » est un groupement peu fiable (les deux odorants suscitent tous les deux plus de jugements positifs que de jugements négatifs). Dans « anoms-é-00 », « lavande, violette » est un groupement fiable (les deux odorants suscitent tous les deux plus de jugements positifs que de jugements négatifs). Dans « snoms-é-00 », « eucalyptus, menthe » est un groupement peu fiable (les deux odorants suscitent tous les deux plus de jugements négatifs que de jugements positifs).

Quand l'un des odorants suscite plus de jugements positifs et l'autre plus de jugements négatifs, on observe la formation d'un groupement fiable, « eucalyptus, menthe » dans « anoms-p-01 » : l'eucalyptus est jugé plus positif que négatif, la menthe est jugée plus négative que positive. On n'observe pas les autres cas de figure. Mais on peut supposer que les odorants testés et les restrictions que nous nous sommes fixées ne permettent pas d'en faire état.

<sup>49</sup> Dans « anoms-é-00 », « eucalyptus, menthe » est un groupement fiable. Il est décrit avec autant de termes positifs que de termes négatifs. On retrouve ici le fait que des sujets différents regroupent des odorants pour des raisons opposées.

<sup>50</sup> Ils peuvent être opposés tout en partageant la même valeur : par exemple, le sujet 9 de « anoms-p-01 » décrit le groupement impliquant la vanille comme *fruité, agréable* et le groupement impliquant la cannelle comme *épice, agréable*.

Ils peuvent être opposés sur des critères différents, impliquant ou pas le jugement hédonique : par exemple, le sujet 8 de « anoms-p-01 » décrit le groupement impliquant la vanille comme *nourriture, papilles gustatives* et le groupement impliquant la cannelle comme *odeur forte et dérangeante* ; le sujet 10 de « anoms-p-01 » décrit le groupement impliquant la vanille comme *se mange* et le groupement impliquant la cannelle comme *encens, pots pourris*.

<b>anoms-p-01</b>	<b>V</b>	<b>N</b>	<b>Co</b>	<b>M</b>	<b>Ch</b>	<b>P</b>	<b>anoms-p-01</b>	<b>V</b>	<b>N</b>	<b>Co</b>	<b>M</b>	<b>Ch</b>	<b>P</b>
<b>vanille et cannelle</b>	oui	oui					<b>pomme et mûre</b>	oui	oui			oui	
<b>vanille sans cannelle</b>	oui	oui	oui		oui		<b>pomme sans mûre</b>	oui	oui		oui	oui	
<b>cannelle sans vanille</b>	oui	oui				oui	<b>mûre sans pomme</b>	oui	oui	oui			
<b>lavande et violette</b>	oui					oui							
<b>lavande sans violette</b>	oui	oui		oui									
<b>violette sans lavande</b>	oui	oui	oui		oui								

<b>snoms-é-00</b>	<b>V</b>	<b>N</b>	<b>Co</b>	<b>M</b>	<b>P</b>	<b>Ch</b>	<b>snoms-é-00</b>	<b>V</b>	<b>N</b>	<b>Co</b>	<b>M</b>	<b>P</b>	<b>Ch</b>
<b>vanille et cannelle</b>	oui	oui					<b>violette et mûre</b>	oui		oui			
<b>vanille sans cannelle</b>		oui	oui	oui			<b>violette sans mûre</b>	oui	oui	oui			oui
<b>cannelle sans vanille</b>	oui	oui		oui			<b>mûre sans violette</b>		oui		oui	oui	oui
<b>eucalyptus et menthe</b>		oui	oui	oui			<b>lavande et herbe</b>	oui					
<b>eucalyptus sans menthe</b>	oui	oui		oui		oui	<b>lavande sans herbe</b>	oui	oui		oui		
<b>menthe sans eucalyptus</b>	oui	oui			oui		<b>herbe sans lavande</b>	oui	oui				
<b>pomme et chewing-gum</b>		oui											
<b>pomme sans chewing-gum</b>	oui	oui											
<b>chewing-gum sans pomme</b>	oui	oui	oui										

Tableau 23 : répartition référentielle pour les odorants des groupements peu fiables

On voit que les odorants étudiés sont le plus souvent spécifiés pour au moins deux domaines de connaissances, domaines que l'on retrouve toujours au moins une fois pour un autre groupement.

Cette diversité de spécifications peut alors constituer une autre des raisons pour lesquelles on obtient une aussi grande variation dans les résultats des expériences : la saisie des odorants à travers des domaines d'activités multiples favorise la diversité des regroupements. Toutefois, si l'on se rappelle que le groupement fiable « orange, citron », dans « anoms-p-01 », était spécifié pour les domaines végétal (quasiment restreint aux fruits), nourriture, cosmétique, produit d'entretien et chimie (avec en outre le jugement hédonique, l'intensité, la familiarité et la typicité), la formulation de l'hypothèse n'est pas assez précise. On doit y inclure la mention d'odorants ne présentant pas de caractéristiques olfactives suffisantes et/ou celle des odorants qui suscitent des jugements hédoniques opposés. Ces deux propriétés doivent en outre être mises en relation avec les cadres d'identification que se donnent les sujets. En effet, pour eux, il faut, tout au moins dans les expériences « snoms » :

(i) trouver un « nom » à l'odorant ;

(ii) trouver un « nom » présentant les « bonnes » caractéristiques cognitives et linguistiques ;

(iii) établir un classement cohérent.

Or on remarque que, pour beaucoup de sujets, les points (i) et (ii) s'avèrent problématiques, et qu'ils ne donnent pas forcément les clés pour établir le point (iii) (cf. note 29).

Bien que nous ne sachions pas précisément comment se nouent les interactions entre ces divers ordres de contraintes et de principes, il n'en reste pas moins que l'hypothèse d'une organisation catégorielle en prise sur les pratiques et les activités des sujets dessine un cadre général productif : la diversité des domaines de connaissances renvoie à la diversité des domaines dans lesquels les odeurs sont rencontrées, et par suite, à la diversité des pratiques des sujets dans ces mêmes domaines.

### ***Bilan***

Deux critères ont ici été utilisés. Le premier, le nombre d'objets associés, s'est révélé peu productif par rapport à nos données. Le second, l'analyse des verbalisations, a permis de poser différentes hypothèses expliquant pourquoi tel ou tel groupement restait peu fiable. Même si nous nous sommes restreintes à l'étude des groupements peu fiables qui constituent par ailleurs des groupements fiables dans au moins une des autres expériences, cela a permis de mettre à jour un double jeu de contraintes dans l'organisation catégorielle des odeurs, qui rend compte des « attractions » différenciées entre les odeurs/odorants :

(i) le jugement hédonique différencié porté sur un même odorant ;

(ii) les pratiques et leur diversité, qui tendent à saisir l'odeur/odorant testé(e) sous une multiplicité de points de vue.

## **Conclusion**

L'utilisation d'outils permettant de faire émerger des dimensions, des classes, etc. demande de manière générale de connaître précisément les fondements mathématiques sur lesquels ces outils sont fondés, et par suite, les choix, les biais que chacun impose. Nous avons illustré ce point avec une méthode de classification fondée sur les arbres de similarité additifs. L'investigation a porté sur l'existence et la robustesse de catégories d'odeurs.

Comme on pouvait s'y attendre au vu de la littérature, les odeurs ne se laissent pas aisément représenter par une distance d'arbres. La valeur des indicateurs portant sur la topologie dans son ensemble, sans être faible, n'est pas suffisante pour dissiper tous les doutes quant à l'organisation catégorielle produite. Toutefois, si l'on voulait généraliser ce premier résultat, il faudrait étendre l'étude à d'autres odorants, pris dans différentes structures d'échantillons. En effet,



(i) l'échantillon testé a été essentiellement retenu sur la base de la familiarité, propriété qui était certes pertinente par rapport au cadre de travail retenu, mais qui pourrait être conjuguée avec d'autres ;

(ii) comme le groupe des mauvaises odeurs apparaît déterminant dans la structuration de cet échantillon, il serait nécessaire de préciser ce qui se passe sur le versant positif avec d'autres exemples (cf. ci-dessous).

Nous avons montré cependant que l'on pouvait travailler non plus sur la topologie dans son ensemble, mais sur certaines de ses parties. En d'autres termes, si les critères de fiabilité ne permettent pas de valider l'ensemble de l'arborescence obtenue, certaines parties de l'arbre peuvent l'être. Deux sortes de parties ont été explorées :

1/ Les « groupements fiables », isolés sur la base de critères ayant une valeur élevée. Nous avons déterminé leur type et précisé leur organisation catégorielle. Les groupements « fuel, poisson, ail » et « orange, citron » sont de type différent, et l'on peut inférer pour le premier une organisation en termes de ressemblance de famille. Nous avons aussi montré l'importance du jugement hédonique et posé l'hypothèse que, dans son versant négatif, il agissait comme un principe organisateur de niveau de base. Qu'il constitue aussi dans son versant positif un principe organisateur de niveau de base reste une question ouverte. Ce qui semble en revanche assuré est que ce jugement ne présente pas un caractère dimensionnel. Ces différents résultats confirment d'autres expériences menées d'autres points de vue (Rouby et Bensafi 2002 ; Bensafi *et al.* 2002a, b et c).

2/ Les « groupements peu fiables », isolés de manière symétrique aux premiers, sur la base de critères ayant une valeur peu élevée. Dans l'analyse, nous nous sommes limitées aux groupements impliquant uniquement des distances d'arêtes externes et qui constituent un groupement fiable dans au moins une des autres expériences. Cette deuxième condition se révèle particulièrement intéressante, car elle permet d'examiner plus précisément les raisons pour lesquelles la catégorisation est fluctuante d'une expérience à l'autre. Plutôt que de questionner l'existence de telle ou telle catégorie d'odeurs, nous avons étudié ici les principes classificateurs à l'œuvre. De cet examen, il ressort l'importance du jugement hédonique et des domaines d'activités et de pratiques.

Appliquée à la catégorisation des odeurs, domaine encore peu exploré, l'utilisation d'outils constituant des classes structurées est une méthode productive : elle n'impose pas de hiérarchie *a priori* ; elle permet d'explorer plus spécifiquement certaines parties de la représentation ; elle a une forte valeur heuristique notamment quand on analyse les résultats produits avec les différents types de données sources.

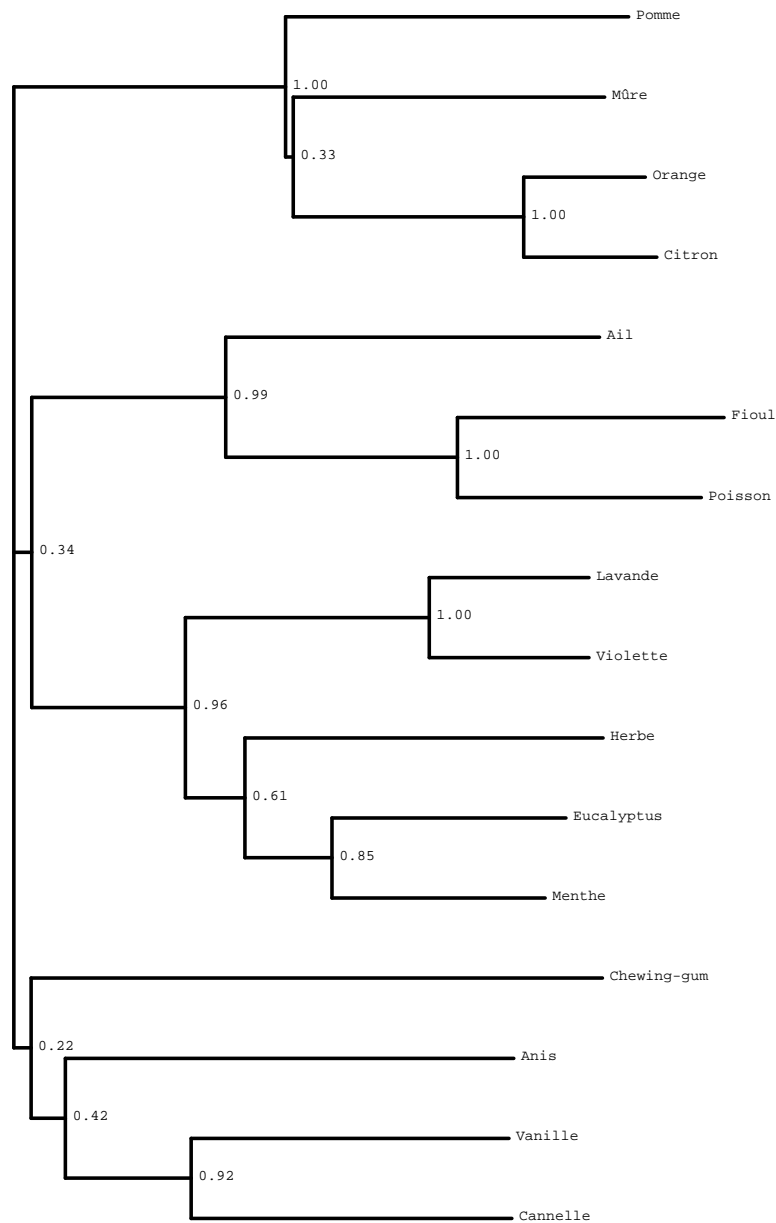
## Références

- Aubry, V., Etiévant, P., Sauvageot, F. & Issanchou, S. (1999) Sensory analysis of Burgundy Pinot noir wine : orthonasal and retronasal profiling, *Journal of Sensory Studies*, 14, 97-117.
- Barthélemy, J.-P. (1991) Similitude, arbres et typicalité, in D. Dubois (éd.), *Sémantique et cognition*. Paris, CNRS Editions, 205-224.
- Barthélemy, J.-P. & Guénoche, A. (1988) *Les arbres et les représentations de proximité*, Paris, Masson.
- Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2001) Réponses émotionnelles aux odeurs : aspects psychologiques et psychophysiologiques, in H. Paugam-Moisy, V. Nyckees & J. Caron-Pargue (éds), *La cognition entre individu et société*, Paris, Hermès Science Publication, 191-202.
- Bensafi, M., Rouby, C., Farget, V., Vigouroux, M. & Holley, A. (2002a) Asymmetry of pleasant vs. unpleasant odor processing during affective judgments in humans, *Neuroscience Letters*, 328, 309-313.
- Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2002b) Influence of affective and cognitive judgments on autonomic parameters during inhalation of pleasant and unpleasant odors in humans, *Neuroscience Letters*, 319, 162-166.
- Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2002c) Autonomic nervous system and odors : the role of pleasantness and arousal, *Chemical Senses*, 27, 703-709.
- Berglund, B., Berglund, U., Engen, T. & Ekman, G. (1973) Multidimensional analysis of twenty-one odors, *Scandinavian Journal of Psychology*, 14, 131-137.
- Boisson, C. (1997) La dénomination des odeurs : variations et régularités linguistiques. *Intellectica*, 24, 1, 29-49.
- Camargo, E. (1996) Valeurs médiatives en Caxinaua, in Z. Guentchéva (éd.), *L'énonciation médiatisée*, Louvain-Paris, Peeters, 271-283.
- Carrasco, M. & Ridout, J.-B. (1993) Olfactory perception and olfactory imagery : a multidimensional analysis, *Journal Experimental Psychology: Human Perception and Performance*, 19, 287-301.
- Chastrette, M. (2002) Classification of odors and structure-odor relationships, in C. Rouby, B. Schaal, D. Dubois, R. Gervais & A. Holley (éds), *Olfaction, Taste and Cognition*. Cambridge, Cambridge University Press, 100-116.
- Classen, C. (1993) *Worlds of Senses*, London-New-York, Routledge.
- Corbin, D. & Temple, M. (1994) Le monde des mots et des sens construits, *Cahiers de Lexicologie*, 65, vol. 2, 5-28.
- David, S. (2002) Linguistic expressions of odors in French, in C. Rouby, B. Schaal, D. Dubois, R. Gervais & A. Holley (éds), *Olfaction, Taste and Cognition*, Cambridge, Cambridge University Press, 82-99.
- David, S. (2000) Certitudes et incertitudes dans les domaines olfactif, gustatif et auditif, *Cahiers du LCPE*, 4, 77-108.
- David, S. (1997) Représentations sensorielles et marques de la personne : contrastes entre olfaction et audition, in D. Dubois (éd.), *Catégorisation et cognition. De la perception au discours*, Paris, Kimé, 211-242.
- David, S., Dubois, D. & Rouby, C. (2000) Lexical devices and the construction of « objects » : a comparison between sensory modes. In A. Melby et A. R. Lommel (éds), *LACUS Forum XXVI*, Fullerton (Californie) : The Linguistic Association of Canada, 225-235.
- David, S., Dubois, D., Rouby, C. & Schaal, B. (1997) L'expression des odeurs en français : analyse lexicale et représentation cognitive, *Intellectica*, 24, 1, 51-83.
- Distel, H. et Hudson, R. (2001) Judgement of odor intensity is influenced by subjects' knowledge of the odor source, *Chemical Senses*, 20, 247-251.

- Dubois, D. (éd.) (1991) *Sémantique et cognition*. Paris, Éditions du CNRS.
- Dubois, D. (1993) Lexique et catégories naturelles : représentations ou connaissances ? *Cahiers de praxématique*, 21, 105-124.
- Dubois, D. (éd.) (1997) *Catégorisation et cognition. De la perception au discours*, Paris, Kimé.
- Dubois, D. (2000) Categorization as acts of meaning : the case of categories in olfaction and audition, *Cognitive Science Quarterly*, 1, 35-68.
- Dubois, D. & Grinevald, C. (2000a) Denominations of colors in practices, in A. Melby et A. R. Lommel (éds), *LACUS Forum XXVI*. Fullerton (Californie), The Linguistic Association of Canada, 237-245.
- Dubois, D. & Rouby, C. (1997) Une approche de l'olfaction : du linguistique au neuronal. *Intellectica*, 24, 1, 9-20.
- Dubois, D. & Rouby, C. (2002) Names and categories for odors : the veridical label, in C. Rouby, B. Schaal, D. Dubois, R. Gervais, A. Holley (éds), *Olfaction, Taste and Cognition*, Cambridge, Cambridge University Press, 47-66.
- Dubois D., Rouby, C. & Sicard, G. (1997) Catégories olfactives et sensorialités : de l'espace visuel à l'espace olfactif, *Enfance*, n° spécial, *L'olfaction chez l'enfant : perspectives croisées*, 1, 141-151.
- Engen, T. (1987) Remembering odors and their names, *American Scientist*, 75, 497-503.
- Gilbert, A. N. & Kemp, S. E. (1996) Odor perception phenotypes : multiple, specific hyposmias to musks, *Chemical Senses*, 20, 411-416.
- Godard, D. & Jayez, J. (1996) Types nominaux et anaphores : le cas des objets et des événements. *Chronos*, 1, 41-58.
- Godinot, N. (1994) *Perception et catégorisation des odeurs par l'Homme*, DEA de Neurosciences, Université Claude Bernard Lyon I (<http://olfac.univ-lyon1.fr>).
- Godinot, N. (1999) *Contribution à l'étude de la perception olfactive : qualité des odeurs et mélanges de composés odorants*, Thèse de doctorat, Université Claude Bernard Lyon 1.
- Godinot, N., David, S., Dubois, D., Bouchet-Rallet, E. & Rouby, C. (1999) Comparative study of (19 synthetic aromas of Lemon, Orange and Mandarin and their respective mixtures: from chemical compositions to linguistic representations, *Chemical Senses*, 24, 1, 71.
- Guénoche, A. (1999) Catégories, ms.
- Guénoche, A. & Garreta, H. (2001) Can we have confidence in a tree representation ? *Proceedings of JOBIM' 2000, Lecture Notes in Computer Sciences*, 2066, 43-53.
- Issanchou, S., Schlich, P. & Guichard, E. (1989) Profils d'odeur des constituants de l'arôme de l'abricot : description par analyse des correspondances, *Sciences des Aliments*, 9, 351-370.
- Kleiber, G. (1984) Dénominations et relations dénominatives, *Langages*, 76, 77-94.
- Kleiber, G. (1990) *La sémantique du prototype. Catégories et sens lexical*, Paris, PUF.
- Lakoff, G. (1987) *Women, fire and dangerous things*, Chicago, University of Chicago Press.
- Mennecier, Ph. & Robbe, B. (1996) La médiation dans le discours des Inuits, in Z. Guentchéva (éd.), *L'énonciation médiatisée*, Louvain-Paris, Peeters, 233-247.
- Mouéle, M. (1997) L'apprentissage des odeurs chez les Waanzi, note de recherche, *Enfance*, 1, 209-222.
- Putnam, H. (1975) La sémantique est-elle possible ? In J. Chaurand et F. Mazière (éds), *La définition*, Paris : Larousse, 292-304, trad. J.-M. Marandin.
- Rosch, E., Mervis, C., Gray, W., Johnson, D. & Boyes-Braem, P. (1976) Basic objects in natural categories, *Cognitive Psychology*, 8, 382-439.
- Rosch, E. (1978) Principles of categorization, in E. Rosch et B. Lloyd (éds), *Cognition and Categorization*. Hillsdale : Erlbaum, 27-48.
- Rouby, C. & Bensafi, M. (2002) Is there a hedonic dimension to odors ? In C. Rouby, B. Schaal, D. Dubois, R. Gervais, R. & A. Holley (éds), *Olfaction, Taste and Cognition*. Cambridge : Cambridge University Press, 140-159.
- Rouby, C. & Sicard, G. (1997) Des catégories d'odeurs ? In D. Dubois (éd.), *Catégorisation et Cognition. De la perception au discours*, Paris, Kimé, 59-81.

- Saitou, N. & Nei, M. (1987) The neighbor-joining method : a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 4, 406-425.
- Sattath, S. & Tversky, A. (1977) Additive similarity trees, *Psychometrika*, 42, 319-345.
- Schaal, B., Rouby, C., Marlier, L., Soussignan, R. & Tremblay, R. E. (1998) Variabilité et universaux de l'espace perçu des odeurs : approches inter-culturelles de l'hédonisme olfactif, in R. Dulau & J.-R. Pitte (éds), *Géographie des odeurs*, Paris, L'Harmattan, 25-47.
- Schiffman, S. S. (1974) Physicochemical correlates of olfactory quality, *Science*, 185, 112-117.
- Schiffman, S. S., Robinson, D. & Erickson, R. P. (1977) Multidimensional scaling of odorants : examination of psychological and physiological dimensions, *Chemical Senses*, 2, 375-390.
- Urdapilletta, I., Ton Nu, Ch., Saint-Denis, C. & Huon de Kermadec, F. (2001) *Traité d'évaluation sensorielle. Aspects cognitifs et métrologiques*, Paris, Dunod.
- Logiciels disponibles aux adresses suivantes :
- Addtree : <http://www.fas.umontreal.ca/biol/legendre>
- Treeview : <http://taxonomy.zoology.gla.ac.uk/rod/rod.html>

**Annexe : représentation construite par Treeview pour  
« étiqu-p-01 »**



1

Le nombre indiqué à chaque nœud exprime le taux de quadruplets supportés par l'arête interne. Le trait indiqué sous 1 exprime l'échelle.



# Méthodes de partition centrale appliquées à l'étude de catégories cognitives<sup>1</sup>

**Thierry Brenac**

(Institut national de recherche sur les transports et leur sécurité, INRETS)

## Introduction

Dans le domaine de la sécurité de systèmes dynamiques comme les systèmes de transports, l'étude des représentations catégorielles ou prototypiques (concepts issus de la psychologie cognitive : voir notamment Rosch, 1973, 1978a, Dubois, 1991) en mémoire permanente des opérateurs humains revêt une importance particulière, à la fois du point de vue de la compréhension des mécanismes accidentels (Van Elslande, 2001, Van Elslande et Alberton, 1997), du point de vue de l'analyse des fonctionnements des systèmes et pour l'amélioration de leur conception (voir en particulier Fleury *et al.*, 1991). La régulation de systèmes évoluant avec une grande rapidité nécessite en effet, au-delà de la mise en œuvre d'automatismes, le recours à des procédures de traitement de l'information à faible coût cognitif, mobilisant des représentations catégorielles, des prototypes de situations déjà acquis en mémoire permanente, de préférence à la mise en œuvre plus coûteuse de fonctionnements de type « résolution de problème » (on se réfère ici au modèle de fonctionnement cognitif de l'opérateur de Rasmussen : Rasmussen, 1974, Leplat, 1985). Le concept de représentation prototypique est également pertinent en matière d'organisation et de structuration de connaissances expertes (Cantor *et al.*, 1980), et est utilisé notamment, pour ce qui concerne la sécurité des transports, pour l'explicitation de connaissances sur les accidents de la circulation routière (Fleury et Brenac, 2001, Brenac et Fleury, 1999, Després, 1998). Un ensemble de recherches, utilisant les concepts de catégorie cognitive et de prototype pour des applications au domaine de la sécurité des transports, a donc été conduit à l'INRETS depuis une dizaine d'années, souvent en association

---

<sup>1</sup> Les contenus de cet article ont fait l'objet par ailleurs d'une publication plus complète sous forme d'un rapport de recherche (Brenac, 2002).

avec le LCPE (voir notamment Fleury *et al.*, 1993, Dubois et Fleury, 1994, Dubois *et al.*, 1993).

La mise en évidence expérimentale de telles catégories cognitives peut s'appuyer notamment sur des classements d'objets — le terme « objet » étant pris ici dans un sens très large (objets simples, scènes, événements, présentés sous forme de photographies, sous forme textuelle, *etc.*) — réalisés par des sujets d'expérimentation. L'analyse des données expérimentales ainsi constituées comporte un certain nombre de difficultés méthodologiques, concernant notamment le passage des regroupements en classes opérés par les différents sujets à une classification ou partition centrale susceptible de rendre compte de représentations catégorielles générales. Dans le cadre d'une recherche en cours sur les catégories des conducteurs concernant les voiries urbaines en période nocturne, conduite par M.-C. Montel, différentes questions méthodologiques relatives à ce type de démarche ont été traitées. En particulier, les potentialités des méthodes de partition centrale utilisant la programmation linéaire (Marcotorchino et Michaud, 1982), très coûteuses en temps de calcul, ont été ré-examinées, eu égard à l'évolution des capacités des micro-ordinateurs.

Après une brève présentation des différentes méthodes envisageables pour passer des partitions individuelles à des résultats plus généraux, et une discussion sur leur intérêt pour le type de recherche que nous venons d'évoquer, cet article traite principalement d'une méthode de partition centrale par la règle de majorité sous contraintes de Condorcet, reposant sur une résolution exacte par programmation linéaire bivalente, inspirée des principes décrits par Marcotorchino et Michaud (1982). Quelques autres aspects de l'exploitation des données expérimentales produites dans de telles études sont également développés. Enfin, quelques résultats de la recherche évoquée plus haut sur les catégories de voies urbaines sont exposés pour illustrer cet ensemble d'éléments méthodologiques.



## 1. Méthodes pour l'obtention d'une classification ou d'une partition centrale à partir d'un ensemble de partitions

Nous ne ferons pas ici une revue exhaustive et détaillée, mais nous présenterons brièvement les principes généraux des méthodes qui ont été mises en œuvre pour traiter cette question dans les recherches françaises sur les représentations catégorielles, ou dans le domaine de l'agrégation de similarités.

### Méthodes visant une représentation arborée ou une classification hiérarchique

Dans cette famille de méthodes, celles qui ont été utilisées sont principalement la construction d'arbres valués à distance additive (voir : Barthélémy et Guénoche, 1988), et la classification ascendante hiérarchique (CAH) (voir : par exemple Fleury *et al.*, 1991, Xu, 1989).

Soit un ensemble de  $n$  objets à classer, numérotés de 1 à  $n$  et que nous identifierons par leur indice  $i$  (et non par une notation de type  $x_i$ ) pour simplifier les notations. Des  $p$  partitions de cet ensemble réalisées par les  $p$  sujets, on tire d'abord un tableau de proximités, ou de dissimilarités entre les objets classés (photographies de sites routiers par exemple). Chaque partition initiale, définie par un sujet de l'expérimentation, peut être décrite par une table de comparaison par paires dont chaque terme  $comp(i, j)$  vaut 1 si les objets  $i$  et  $j$  ont été groupés dans la même classe et 0 dans le cas contraire. La matrice somme de ces tables pour l'ensemble de ces sujets, dont chaque terme  $c_1(i, j)$  représente le nombre de sujets ayant classé  $i$  et  $j$  ensemble, constitue le tableau de proximités sur lequel opèrent les méthodes qui suivent.

#### *Arbres valués à distance additive*

Les méthodes de représentation des proximités par des arbres à distance additive (*cf.* Barthélémy et Guénoche, 1988), d'abord développées en vue d'applications à des questions d'évolution ou de filiation (en biologie, mais aussi dans le domaine de la paléographie, de l'archéologie, de la filiation de manuscrits ; Buneman, 1971), ou dans le domaine du traitement de l'information (Zaretskii, 1965, Smolenskii, 1969, cités par Barthélémy et Guénoche), ont été ensuite également appliquées à des questions de psychologie cognitive, et en particulier à la représentation de catégories cognitives et de l'organisation hiérarchique des connaissances (voir notamment A. Tversky, 1977, Cunningham, 1978). Ces méthodes sont celles qui ont été le plus fréquemment utilisées dans les recherches françaises sur les représentations catégorielles.

Elles produisent une représentation arborescente des proximités entre objets, sous forme d'un X-arbre valué à distance additive (voir notamment Buneman, 1971 ; Sattah et A. Tversky, 1977 ; Guénoche, 1986, Barthélémy et Luong,

1986 ; Barthélémy et Guénoche, 1988, où l'on trouvera des notes bibliographiques très complètes). Un X-arbre est un arbre dont certains sommets sont réels et correspondent dans notre cas à des objets, d'autres sommets étant latents. Les sommets réels sont les éléments d'un ensemble X, défini une fois pour toutes. Les sommets latents jouent le rôle d'intermédiaires dont le statut peut être différent de celui des objets (genres, familles, ordres, par exemple, dans le domaine de la classification systématique des espèces animales ou végétales), mais ils peuvent dans d'autres cas renvoyer à des objets inconnus (« chaînons manquants » en phylogénétique ou dans d'autres domaines comme la filiation de manuscrits) (Barthélémy et Guénoche, 1988). Un tel arbre est valué si à chacune de ses arêtes est associée une valeur positive. Un X-arbre valué à distance additive est un X-arbre valué pour lequel ces valeurs représentent des distances entre sommets telles que la distance entre deux sommets quelconques de l'arbre est égale à la somme des valeurs affectées aux arêtes qui permettent de les joindre. En pratique, on dispose d'un indice de proximité entre les objets à classer et la construction de l'arbre permet, en jouant sur l'ajout et le positionnement de sommets latents (nœuds intermédiaires), d'obtenir une distance additive constituant une modélisation (ou approximation, selon le point de vue adopté) de cet indice de proximité.

On peut déduire du tableau de proximités évoqué plus haut un indice  $dis(i,j) = p - c_1(i,j)$ , qui est un indice de dissimilarité (ou indice d'écart selon la terminologie de Barthélémy et Guénoche, 1988) : il est symétrique, positif, et tel que  $dis(i,i) = 0$  pour tout  $i$ . On peut montrer que cet indice vérifie l'inégalité triangulaire (et constitue donc un « écart » selon cette même terminologie).

Cet indice ne vérifie pas nécessairement, dans le cas général, la « condition des quatre points », alors que cette propriété  $[d(x,y) + d(z,t) \leq \text{Max}\{d(x,z) + d(y,t), d(x,t) + d(y,z)\}]$  pour tout quadruplet  $x, y, z, t$  est une propriété des distances ou écarts arborés, représentables sous forme d'arbres valués à distance additive (Barthélémy et Guénoche, 1988). Hormis dans le cas où la configuration des données serait d'emblée en cohérence parfaite avec une structure d'arbre à distance additive, on procède donc en général à la modélisation (ou approximation) de cet indice par une distance ou un écart arboré. Parmi les différentes méthodes permettant d'accéder à cette approximation, ce sont les méthodes reposant sur le concept de score (Sattath et A. Tversky, 1977, Barthélémy et Luong, 1986), qui ont été généralement utilisées dans les recherches françaises sur les représentations catégorielles : l'indice de dissimilarité sert d'abord à construire des « scores », qui traduisent la « force » d'un voisinage entre deux sommets en tenant compte également des paires d'éléments opposées à ces deux sommets, et permettent de construire itérativement un arbre valué à distance additive. Nous ne détaillerons pas ici ces méthodes, sur lesquelles il existe déjà des publications très complètes (le lecteur pourra se reporter à la référence : Barthélémy et Guénoche, 1988, et en particulier au chapitre IV).

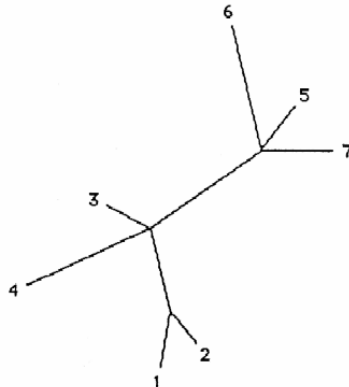


Figure 1. Exemple d'arbre valué à distance additive (source : Barthélémy et Guénoche, 1988). Les sommets numérotés représentent des sommets réels, les sommets intermédiaires, non numérotés, sont des sommets latents.

Les représentations arborées obtenues par ces méthodes permettent notamment de mettre en évidence certaines conclusions en matière de filiation (notions d'« ancêtre commun », de « chaînon manquant »), qui peuvent avoir une grande importance dans certains domaines (phylogénétique, filiation de manuscrits, notamment ; voir par exemple Cavalli-Sforza et Edwards, 1967, Klotz et Blanken, 1981, Buneman, 1971). Concernant l'analyse des représentations catégorielles, la notion de filiation a cependant moins de pertinence, mais la structure arborescente permet de rendre compte de l'organisation hiérarchique des connaissances en mémoire. Un autre avantage des arbres à distance additive est qu'ils permettent une représentation de la typicalité d'un individu (par la distance de ce sommet au nœud auquel se rattachent les individus de la classe), si l'on suppose que les nœuds ou sommets latents représentent le positionnement des prototypes.

### **Méthodes de classification hiérarchique**

Des méthodes de classification hiérarchique ascendante ont été appliquées à ce type de problème, par exemple par Xu (1989), utilisant comme distance entre individus (objets) la distance euclidienne entre lignes du tableau disjonctif complet rendant compte des partitions des différents sujets (chaque sujet est assimilé à une variable qualitative, et chacune de ses classes à une modalité de cette variable) et comme ultramétrie celle de l'indice de Ward ou celle du saut minimal (lien simple).

Ces méthodes, sur lesquelles nous ne nous étendrons pas ici dans la mesure où elles peuvent être considérées comme classiques, produisent une hiérarchie indicée de parties de l'ensemble des objets. Les options et méthodes possibles sont assez diverses, du fait notamment des nombreux choix envisageables concernant la métrique d'origine et l'ultramétrie. Pour des ensembles de données limités (14 objets au plus), une hiérarchie optimale au sens des moindres carrés peut être obtenue (Chandon, Lemaire et Pouget, 1980, cités par Roux,

1998) et permet alors de comparer les résultats obtenus par différentes méthodes de classification hiérarchique (Roux, 1998).

Concernant l'application à la mise en évidence de représentations catégorielles en mémoire permanente, on peut noter que ces méthodes, comme les précédentes, permettent de figurer une organisation hiérarchique des catégories. Les niveaux d'agrégation permettent de rendre compte des distances entre classes. À la différence des arbres à distance additive, les dendrogrammes de classification hiérarchique ne sont pas adaptés à la représentation de la plus ou moins grande typicalité des individus (objets). En effet, si l'on cherche à caractériser la distance entre un individu  $i$  et le prototype auquel il se rattache, sur la base du dendrogramme obtenu, on est amené à considérer la distance entre cet individu et sa classe d'appartenance  $C$ , en la définissant par la distance ultramétrique  $\mathbb{D}(\{i\}, C)$  où  $\mathbb{D}$  est l'ultramétrie choisie pour la construction de la classification. Or cette distance est identique pour tous les individus de la classe, du fait des propriétés des ultramétries. Le concept de représentation prototypique implique que les différents exemplaires « sont identifiés comme membres catégoriels, non pas parce qu'ils possèdent les attributs « nécessaires et suffisants », mais en raison de leur « ressemblance » au « prototype » (Dubois et Mazet, 1988) ; cette ressemblance, plus ou moins marquée, devrait se traduire par une plus ou moins grande distance des individus au prototype (contrairement à une catégorisation analytique où tous les individus sont dans une position identique, du point de vue des attributs nécessaires et suffisants, vis-à-vis de leur catégorie).

### Méthodes de partitionnement automatique

Des méthodes de type « nuées dynamiques » ont été utilisées, par exemple par Xu (1989, méthode des centres mobiles), soit en partant du tableau d'indices de dissimilarité  $dis(i,j)$  évoqué plus haut, soit en utilisant une distance (distance du  $\chi^2$  dans les travaux de Xu) entre objets déduite du tableau disjonctif complet mentionné au point précédent. Ces méthodes conduisent à une partition de l'ensemble des objets.

Dans ces méthodes, le nombre de classes doit être fixé *a priori*. D'autre part, les résultats dépendent de l'initialisation (classes de départ). Il est possible pour tenter d'obtenir des classes plus stables, de procéder à plusieurs passages de l'algorithme en faisant varier les conditions initiales, et de rechercher une partition centrale à partir des partitions obtenues en utilisant une heuristique de recherche de partition centrale telle que celle proposée par Celeux (1984).

Comme les méthodes de partition centrale évoquées au point suivant, de telles méthodes ne sont pas adaptées à la représentation de la hiérarchie des connaissances en mémoire, mais sont plus particulièrement intéressantes pour les recherches portant sur un niveau de base des catégories.

Cependant, dans ces techniques de partitionnement automatique, le résultat ne constitue *a priori* qu'un optimum local.

### Méthodes de partition centrale

Ces méthodes visent à déduire directement une partition centrale, constituant une sorte de compromis entre les partitions produites par les différents sujets. On peut notamment distinguer parmi ces méthodes : les méthodes heuristiques reposant sur l'algorithme des transferts de S. Régnier (proposé dès 1965 ; voir par exemple : Régnier, 1983), et les méthodes reposant sur des algorithmes de résolution exacte par programmation linéaire en nombres entiers (Marcotorchinaud et Michaud, 1979, 1982). Mentionnons également l'heuristique évoquée au paragraphe précédent, qui a été proposée par Celeux (1984) dans le cadre plus particulier des méthodes de partitionnement automatique.

L'objectif commun de ces méthodes est de trouver une partition centrale telle que le nombre moyen de désaccords entre cette partition et les partitions réalisées par les différents sujets soit minimisé (le nombre de « désaccords » entre deux partitions étant le nombre des paires d'objets telles que les deux objets sont classés ensemble dans une de ces partitions et séparés dans l'autre). Cela est équivalent à trouver une partition telle que la distance moyenne entre elle et les partitions réalisées par les sujets soit minimale — si l'on prend pour distance entre deux partitions le cardinal de la différence symétrique de leurs graphes<sup>2</sup>. Ce problème est rendu particulièrement difficile par le fait que le nombre de partitions possibles d'un ensemble de  $n$  éléments (nombre de Bell) croît de façon « astronomique » en fonction de  $n$ .

Les résultats se présentant sous forme d'une série de classes disjointes, ils ne rendent donc pas compte en eux-mêmes des aspects de prototypie. Chaque classe regroupe différents exemplaires plus ou moins proches du prototype, dont la mise en évidence nécessite un examen de la typicalité des exemplaires (voir point 3) et un travail d'interprétation. Les objets les plus éloignés des prototypes, qui ne sont classés avec tout autre objet que par un petit nombre de sujets, tendent à former des classes à un seul élément (voir l'application).

---

<sup>2</sup> Par graphe d'une partition, nous entendons le graphe de la relation d'équivalence  $\mathcal{R}$  qui la définit, ce graphe étant l'ensemble des couples  $(i, j)$  tels que  $i \mathcal{R} j$ , c'est-à-dire  $i$  'dans la même classe que'  $j$ . La différence symétrique des graphes de deux partitions  $P_1$  et  $P_2$  est donc l'ensemble des couples  $(i, j)$  tels que  $i$  et  $j$  sont classés ensemble pour  $P_1$  mais non pour  $P_2$  ou bien tels que  $i$  et  $j$  sont classés ensemble pour  $P_2$  mais non pour  $P_1$ . Le nombre d'éléments (cardinal) de cet ensemble (différence symétrique des deux graphes) est un indice de dissimilarité entre les deux partitions. On montre sans difficultés qu'il vérifie les propriétés d'une distance, en particulier l'inégalité triangulaire.

### ***Algorithme des transferts***

Le principe de l'algorithme des transferts consiste, à partir d'une partition initiale (par exemple celle d'un des sujets), à chercher à l'améliorer en transférant un objet de sa classe vers une autre classe, ou vers une classe nouvelle. Par améliorer, il faut entendre ici réduire la distance moyenne de cette partition avec l'ensemble des partitions de base. Le résultat est un optimum local, c'est-à-dire une partition qui ne peut plus être améliorée par le transfert d'un élément. Cet algorithme, dû à S. Régnier, a fait l'objet ensuite de certaines extensions : par exemple, un algorithme proposé par Petit (1993) conduit à un optimum qui ne peut plus être amélioré par le transfert d'un sous-ensemble d'éléments d'une même classe.

### ***Partition centrale par la règle de majorité sous contraintes de Condorcet, résolution par programmation linéaire***

Les méthodes reposant sur une résolution exacte par programmation linéaire (Marcotorchino et Michaud, 1979, 1982) sont plus récentes mais s'inscrivent cependant dans une tradition ancienne de recherche de choix collectif, d'opinion majoritaire ou consensuelle à partir d'ensembles d'opinions individuelles, si l'on se réfère aux travaux de Condorcet au dix-huitième siècle, s'intéressant notamment au classement par ordre de  $n$  candidats à partir d'ordres de préférence exprimés individuellement par des votants. L'application d'une règle majoritaire n'est cependant pas simple, dans la mesure où l'agrégation d'opinions individuelles peut conduire à des paradoxes : par exemple, il peut y avoir une majorité de votants pour préférer  $i$  à  $j$ , une majorité pour préférer  $j$  à  $k$ , sans qu'il y ait de majorité pour préférer  $i$  à  $k$  (paradoxe de Condorcet). Pour surmonter cette difficulté, Condorcet a proposé une règle qui consiste à retenir, parmi l'ensemble des classements possibles, celui qui recueille le plus de suffrages, c'est-à-dire qui minimise le nombre de désaccords avec les classements des votants (règle d'agrégation à la majorité sous contraintes de Condorcet) (Michaud, 1982, Montjardet, 1990). Ces questions d'agrégation de préférences ont été tardivement rapprochées des problèmes d'agrégation de similarités qui nous concernent ici (Marcotorchino et Michaud, 1982), où un ensemble d'éléments est réparti en classes d'éléments similaires par différents « juges » (ou votants, sujets, *etc.*). Dans ces derniers problèmes, l'application d'une règle majoritaire peut également conduire à des paradoxes : il peut y avoir une majorité de juges pour classer  $i$  et  $j$  ensemble, une majorité pour classer  $j$  et  $k$  ensemble, sans qu'il y ait une majorité pour classer  $i$  et  $k$  ensemble (paradoxe de Poincaré)<sup>3</sup>. De la même façon, une règle

---

<sup>3</sup> Ce paradoxe n'est plus si paradoxal si l'on considère les questions de prototypie : on peut trouver une ressemblance entre un enfant et sa mère, en trouver également entre cet enfant et son père, sans trouver de ressemblance entre le père et la mère. La relation de ressemblance n'est donc pas rigoureusement transitive. Cependant, l'ensemble des effets Poincaré sur un échantillon d'objets classés par différents sujets dépasse souvent largement ce qui pourrait résulter des rattachements multiples d'exemplaires à plusieurs prototypes, ces effets traduisant également des différences interindividuelles concernant les catégories elles-mêmes. Par exemple, dans l'application décrite au point 5, si l'on recherche les couples d'objets dont la co-appartenance n'est remise en cause par aucun effet Poincaré, on constate que seule une dizaine d'objets sur 65,

d'agrégation à la majorité sous contraintes peut être énoncée : la partition, ou relation d'équivalence, à retenir est celle qui, parmi l'ensemble des partitions possibles, recueille le plus de suffrages, c'est-à-dire minimise le nombre de désaccords avec les partitions initiales.

Il s'agit donc d'ajouter des contraintes pour imposer que les relations finales entre les objets expriment bien une partition, et donc satisfassent aux conditions d'une relation d'équivalence (transitivité notamment), tout en minimisant le nombre de désaccords entre la partition centrale et les  $p$  partitions initiales. L'application de la règle d'agrégation à la majorité sous contraintes revient alors à maximiser la quantité :

$$\sum_i \sum_j [2 c_1(i,j) - p] y_{ij} \quad (1)$$

avec  $c_1(i,j)$  = nombre de sujets ayant classé  $i$  et  $j$  ensemble  
 $y_{ij}$  = inconnues à déterminer, valant 1 ou 0 selon que  $i$  et  $j$  sont ou non classés ensemble dans la partition centrale recherchée  
 $p$  = nombre total de sujets (nombre de partitions initiales)

dans le cadre des contraintes suivantes :

$y_{ij} = y_{ji}$	pour tout $i, j$	(symétrie)
$y_{ij} + y_{jk} - y_{ik} \leq 1$	pour tout $i, j, k$	(transitivité)
$y_{ii} = 1$	pour tout $i$	(réflexivité)
$y_{ij} = 0$ ou $1$	pour tout $i, j$	

Cette maximisation est équivalente à la minimisation du nombre moyen de désaccords avec les partitions initiales (ou de la distance moyenne à ces partitions : pour une démonstration, voir par exemple Saporta, 1990). Ce problème de maximisation peut être traité par des techniques de programmation linéaire bivalente, en utilisant les propriétés de la dualité pour surmonter les difficultés liées à la grande taille du problème (de l'ordre de  $n^2$  variables et  $n^3$  contraintes), comme l'ont montré Marcotorchino et Michaud. Nous y reviendrons plus en détail. La solution obtenue (éventuellement multiple) est une solution exacte, mais elle n'est pas nécessairement unique : il peut exister d'autres solutions.

---

constituant deux groupes, est concernée ; la modification du seuil majoritaire ne change pas notablement ces résultats. En tous les cas, le principe même de recherche d'une partition centrale implique la réduction des effets Poincaré.

## Choix de méthodes

Dans les recherches sur les représentations catégorielles, l'intérêt peut se porter sur l'organisation hiérarchique des connaissances en mémoire, mais on peut à l'inverse privilégier l'étude d'un certain niveau de catégorisation, correspondant au niveau le plus pertinent dans un contexte défini.

Plusieurs recherches (voir en particulier Rosch *et al.*, 1976 ; Rosch, 1978*b* ; B. Tversky et Hemenway, 1983) mettent en effet en évidence une préférence des sujets pour un certain niveau de catégorisation, conduisant à distinguer dans l'organisation des connaissances un niveau superordonné, un niveau de base (préféré), et des catégories plus précises ou sous-ordonnées. Par exemple, si l'on reprend les résultats de Tversky et Hemenway concernant les catégories de scènes environnementales dans un contexte « neutre » (non spécifié), *indoor* et *outdoor* représentent des catégories superordonnées, *beach*, *mountains*, *home*, *school*, représentent des catégories du niveau de base, et d'autres catégories comme *lake beach*, *elementary school*, représentent un niveau sous-ordonné. Ce niveau de base correspondrait aux catégories les plus informatives dans la limite d'un certain coût cognitif (ces catégories n'étant pas plus informatives qu'il n'est nécessaire dans le contexte de leur utilisation). Cette notion de niveau préféré est également pertinente dans un contexte spécifique, notamment lorsqu'il s'agit d'un contexte correspondant à une activité particulière (comme la conduite automobile, par exemple — Mazet *et al.*, 1987), où la limitation du coût cognitif, l'optimisation du rapport entre contenu informatif des catégories et lourdeur des traitements cognitifs, ont une grande importance.

Si l'on s'intéresse aux aspects de hiérarchie (ou de filiation, pour d'autres types d'application), il semble naturel de recourir aux méthodes permettant d'obtenir une représentation arborescente des proximités entre objets, comme la méthode des arbres à distance additive, ou les méthodes de classification hiérarchique. Les dispositifs expérimentaux doivent alors également tenir compte de cette orientation.

Mais, comme nous venons de le mentionner, d'autres recherches, et notamment celles portant sur les représentations catégorielles opérant dans des contextes d'activité spécifiés (comme celui de la conduite automobile), s'intéressent davantage au niveau de base des catégories. Les dispositifs expérimentaux mis en œuvre dans ces recherches, où la consigne conduit le sujet à se situer dans un contexte d'activité donné, et où on lui demande d'effectuer une partition d'un ensemble d'objets en classes disjointes, sont plus particulièrement adaptés à la mise en évidence du niveau de base des catégories associé à ce contexte.

Pour de telles recherches, il est bien sûr possible d'utiliser les méthodes qui viennent d'être évoquées (représentation arborée à distance additive ou



classification hiérarchique) en cherchant à déterminer un niveau de l'arbre susceptible de représenter le niveau de base des catégories. Un tel choix présente l'inconvénient que des hypothèses supplémentaires concernant les seuils de « coupure » de l'arbre sont nécessaires (au-delà des choix de métrique, ou équivalents, nécessaires dans toutes les méthodes). Cependant, il faut ici mentionner certains avantages de la représentation par arbre à distance additive. Tout d'abord, cette méthode permet d'appréhender la position des prototypes, et donc d'apprécier la proximité des différents exemplaires au prototype, c'est-à-dire leur typicalité, par la longueur d'arête les séparant du sommet latent représentant par hypothèse ce prototype (dans les méthodes de partitionnement ou de partition centrale, il reste néanmoins possible d'apprécier la typicalité des exemplaires sur la base d'autres indices, comme nous le verrons plus loin). Plus généralement, ce type de représentation arborée, où les exemplaires apparaissent à plus ou moins grande distance des prototypes, est à cet égard plus conforme au concept de représentation catégorielle, reposant — comme nous l'avons précisé plus haut — sur la plus ou moins grande ressemblance des exemplaires aux différents prototypes, et non sur des conditions nécessaires et suffisantes qui conduisent à des classes aux frontières précisément définies. En contrepartie, il semble que certains aspects des représentations par arbre à distance additive puissent être considérés comme paradoxaux dans le domaine de la représentation des connaissances en mémoire : considérons une classe d'objets ; parmi les éléments de cette classe, l'objet le plus typique de cette classe, et le prototype lui-même, y apparaissent paradoxalement comme les éléments les plus proches des autres classes et prototypes. Cela tient notamment au fait que les sommets latents, représentant par hypothèse les prototypes, sont en position d'intermédiaires au sein de tels arbres (pour d'autres applications de telles représentations arborées, en phylogénétique par exemple, ou plus généralement pour toutes les questions de filiation, cette position d'intermédiaire des sommets latents est au contraire particulièrement adaptée).

Un autre choix possible consiste à utiliser les méthodes visant à obtenir une partition, permettant ainsi un accès plus simple à un niveau déterminé de catégorisation. L'avantage de ces méthodes est que la partition obtenue, rendant compte de catégories constituant alors une sorte de compromis entre les catégories mises en évidence par les différents sujets, se situe, *a priori*, au même niveau de catégorisation, c'est-à-dire au niveau de base dans le cas des dispositifs expérimentaux que nous considérons ici.

Parmi les méthodes visant à obtenir une partition, les méthodes de partitionnement automatique évoquées plus haut présentent différents inconvénients : nécessité de fixer au préalable le nombre de classes, résultats dépendant des conditions d'initialisation de l'algorithme. Si plusieurs passages de l'algorithme permettent d'ajuster le nombre de classes et de construire à partir de ces résultats une partition centrale aux classes plus stables, il n'est pas certain cependant que le

résultat final soit une solution optimale. Cette dernière remarque peut être faite également au sujet des heuristiques visant à obtenir une partition centrale à partir des partitions initiales : le résultat n'est *a priori* qu'un optimum local.

Les méthodes de partition centrale reposant sur une résolution par programmation linéaire bivalente conduisent à un optimum exact, même s'il n'est pas nécessairement unique. Cet optimum correspond à une partition minimisant le nombre moyen de désaccords avec les partitions initiales, ce qui revient à minimiser la distance moyenne avec ces partitions. Les choix de métrique se limitent à celui de la distance entre partitions. Une telle méthode, malgré l'inconvénient restant de la non unicité possible de la solution, semble donc une des mieux adaptées pour les applications portant sur un niveau de base des catégories.

En complément de cette méthode, on peut aussi envisager d'utiliser l'algorithme des transferts pour voir si des optimums *a priori* locaux obtenus par cet algorithme pour différentes conditions d'initialisation ne sont pas en fait des solutions optimales concurrentes de celle obtenue par optimisation.

## 2. Partition centrale par la règle de majorité sous contraintes, méthode de résolution exacte par programmation linéaire en nombres entiers

Cette méthode est due à Marcotorchino et Michaud (1982). Le problème de partition centrale par la règle de la majorité sous contraintes, comme nous l'avons vu au point 1, peut être posé en termes de maximisation :

$$\text{Max } \sum_i \sum_j [2 c_1(i,j) - p] y_{ij} \quad (1)$$

avec  $c_1(i,j)$  = nombre de sujets ayant classé  $i$  et  $j$  ensemble  
 $y_{ij}$  = inconnues à déterminer, valant 1 ou 0 selon que  $i$  et  $j$  sont ou non classés ensemble dans la partition centrale recherchée  
 $p$  = nombre total de sujets (nombre de partitions initiales)

sous les contraintes :

$y_{ij} = y_{ji}$	pour tout $i, j$	(symétrie)
$y_{ij} + y_{jk} - y_{ik} \leq 1$	pour tout $i, j, k$	(transitivité)
$y_{ii} = 1$	pour tout $i$	(réflexivité)
$y_{ij} = 0$ ou $1$	pour tout $i, j$	

Pour traiter ce problème par les techniques classiques d'optimisation en programmation linéaire, Marcotorchino et Michaud (1982) proposent de résoudre d'abord ce système en relâchant dans un premier temps la dernière contrainte pour

la mettre sous la forme  $0 \leq y_{ij} \leq 1$ , des méthodes de coupe pouvant être utilisées ultérieurement pour obtenir des solutions bivalentes (en 0, 1).

Si l'on pose  $c_{ij} = 2 c_1(i,j) - p$ , ce problème s'écrit alors :

$$\begin{aligned} \text{Max} \quad & \sum_i \sum_j c_{ij} y_{ij} & (2) \\ \text{avec} \quad & y_{ij} = y_{ji} & \text{pour tout } i, j \\ & y_{ij} + y_{jk} - y_{ik} \leq 1 & \text{pour tout } i, j, k \\ & y_{ii} = 1 & \text{pour tout } i \\ & 0 \leq y_{ij} \leq 1 & \text{pour tout } i, j \end{aligned}$$

En pratique, dans les nombreuses applications traitées par Marcotorchino et Michaud, comme dans celles que nous avons étudiées, on obtient les solutions bivalentes à ce dernier problème dès cette première étape et il est inutile de mettre en œuvre des méthodes de coupe.

Sachant que  $y_{ji} = y_{ij}$  et que  $c_{ji} = c_{ij}$ ,

$$\text{Max} \sum_i \sum_j c_{ij} y_{ij} \text{ est équivalent à } \text{Max} \left[ \sum_i \sum_{j/i < j} 2 c_{ij} y_{ij} + \sum_i c_{ii} y_{ii} \right]$$

ce qui peut s'écrire, puisque  $c_{ii} = 2 c_1(i,i) - p = 2 p - p = p$  et  $y_{ii} = 1$  (réflexivité) :

$$\text{Max} \left[ \sum_i \sum_{j/i < j} 2 c_{ij} y_{ij} + p n \right]$$

$$\text{cette maximisation étant équivalente à } \text{Max} \sum_i \sum_{j/i < j} c_{ij} y_{ij}$$

Les variables  $y_{ji}$  telles que  $i < j$  n'apparaissant plus dans la fonction objectif (à maximiser), les contraintes  $y_{ij} - y_{ji} = 0$  n'ont alors plus à être prises en compte dans l'optimisation. De même, les variables  $y_{ii}$  n'apparaissent plus, leur valeur étant d'ailleurs connue, et les contraintes de réflexivité peuvent être écartées. Le problème, mis sous une « forme inégalité » conventionnelle, peut alors s'écrire :

$$\text{Max} \sum_i \sum_{j/i < j} c_{ij} y_{ij} \quad (3)$$

$$\begin{aligned} \text{avec} \quad & y_{ij} + y_{jk} - y_{ik} \leq 1 & \text{pour tout } i, j, k \text{ tel que } i < j < k \\ & -y_{ij} + y_{jk} + y_{ik} \leq 1 & \text{pour tout } i, j, k \text{ tel que } i < j < k \\ & y_{ij} - y_{jk} + y_{ik} \leq 1 & \text{pour tout } i, j, k \text{ tel que } i < j < k \\ & y_{ij} \leq 1 & \text{pour tout } i, j \text{ tel que } i < j \\ & -y_{ij} \leq 0 & \text{pour tout } i, j \text{ tel que } i < j \end{aligned}$$

La résolution directe de ce problème, qui présente un nombre de variables de l'ordre de  $n^2$  et un nombre de contraintes de l'ordre de  $n^3$ , n'est pas recommandable dans la mesure où les logiciels d'optimisation le mettraient automatiquement sous forme standard par l'adjonction de variables d'écart, augmentant très considérablement la taille du problème, qui devient alors pratiquement impossible à résoudre sauf pour des faibles valeurs de  $n$ . Aussi la méthode proposée par Marcotorchino et Michaud consiste-t-elle à résoudre le problème dual, qui se présente initialement sous forme standard, et qui comporte un nombre de variables de l'ordre de  $n^3$  et un nombre de contraintes de l'ordre de  $n^2$ .

Le problème dual du problème ci-dessus (primal) peut être formulé comme suit, en utilisant les propriétés de la dualité (voir par exemple : Baillargeon, 1996), si l'on associe les variables duales  $u'_{ijk}$ ,  $u''_{ijk}$ ,  $u'''_{ijk}$ ,  $u'_{ij}$  et  $u''_{ij}$  respectivement aux contraintes  $y_{ij} + y_{jk} - y_{ik} \leq 1$  ( $i < j < k$ ),  $-y_{ij} + y_{jk} + y_{ik} \leq 1$  ( $i < j < k$ ),  $y_{ij} - y_{jk} + y_{ik} \leq 1$  ( $i < j < k$ ),  $y_{ij} \leq 1$  ( $i < j$ ) et  $-y_{ij} \leq 0$  ( $i < j$ ) :

$$\text{Min} \left[ \sum_{\substack{i,j,k \\ i < j < k}} u'_{ijk} + \sum_{\substack{i,j,k \\ i < j < k}} u''_{ijk} + \sum_{\substack{i,j,k \\ i < j < k}} u'''_{ijk} + \sum_{\substack{i,j \\ i < j}} u'_{ij} \right] \quad (4)$$

avec les contraintes :

$$\sum_{\substack{l \\ j < l}} (u'_{ijl} - u''_{ijl} + u'''_{ijl}) + \sum_{\substack{l \\ l < i}} (u'_{lij} + u''_{lij} - u'''_{lij}) + \sum_{\substack{l \\ i < l < j}} (-u'_{ilj} + u''_{ilj} + u'''_{ilj}) + u'_{ij} - u''_{ij} = c_{ij}$$

et avec des conditions de positivité sur toutes les variables duales.

Ce problème se présente, comme nous l'avons mentionné, sous forme standard, et comporte  $n^2(n-1)/2$  variables et  $n(n-1)/2$  contraintes. Il s'agit donc généralement d'un problème de grande taille (à titre d'exemple, pour une application concernant  $n = 65$  objets, il y a 2080 contraintes, 135200 variables, et la matrice des contraintes comporte plus de 281 millions d'éléments dont 397000 valeurs non nulles), qui peut cependant être résolu sur des micro-ordinateurs actuels, malgré des temps de calculs assez longs (surtout pour la génération de la matrice des contraintes, dont le temps de calcul est proportionnel à  $n^5$ ). Les programmes que nous avons développés et utilisés sont présentés dans la référence : Brenac, 2002.

Les solutions au problème primal se déduisent des solutions du problème dual en utilisant les conditions de complémentarité (voir par exemple Baillargeon, 1996). Le dual étant sous forme standard et non sous forme canonique, on utilise la condition de complémentarité suivante : si la  $i^{\text{ème}}$  variable du dual est non nulle, alors la  $i^{\text{ème}}$  contrainte du primal est saturée. L'utilisation des  $n(n-1)$  dernières variables du dual est en général suffisante pour reconstituer les solutions du

primal, la saturation de l'une des  $n(n-1)$  dernières contraintes du primal donnant directement  $y_{ij} = 1$  ou  $y_{ij} = 0$ . Dans certains cas, certaines variables du primal restent indéterminées à ce stade et il faut examiner les autres conditions de complémentarité. Si des indéterminations subsistent encore cela peut correspondre à des solutions multiples au primal. L'énumération complète des différentes solutions bivalentes compatibles avec la solution duale et les conditions de complémentarité, et la vérification de leur optimalité, sont alors en général possibles (du moins sur les problèmes que nous avons étudiés). Nous avons été confrontés à une telle situation sur certains jeux de données (voir l'application); nous avons noté d'autre part que le cas de solutions multiples paraît plus fréquent lorsque le nombre de sujets est pair.

Enfin, à partir de la solution optimale au dual, on peut obtenir sans difficulté la distance moyenne de la partition centrale obtenue (correspondant à la solution du primal) aux partitions initiales : en effet, en application du théorème de la dualité de la programmation linéaire, la valeur prise à l'optimum par la fonction objectif du primal est égale à la valeur prise à l'optimum par la fonction objectif du dual. D'autre part, la distance moyenne (minimisée) de la partition centrale aux partitions initiales s'écrit (voir par exemple Saporta, 1990):

$$D_{moy\ min} = (1/p) \times [ \sum_i \sum_j c_1(i,j) - \text{Max} ( \sum_i \sum_j c_{ij} y_{ij} ) ] \quad (5)$$

$$\text{et} \quad \text{Max} ( \sum_i \sum_j c_{ij} y_{ij} ) = 2 [ \text{Max} ( \sum_i \sum_{j/i < j} c_{ij} y_{ij} ) ] + p n$$

### 3. Analyse de la partition centrale

L'obtention d'une (ou plusieurs) partition centrale, permet, par l'interprétation des classes d'objets ainsi rassemblés, conduite notamment sur la base de l'expertise des expérimentateurs et de l'analyse des entretiens avec les sujets, d'identifier les principaux traits des représentations catégorielles. D'autres questions peuvent être également abordées concernant par exemple la proximité entre certaines catégories, ou la situation des différents exemplaires relativement aux catégories (plus ou moins grande typicalité). Nous nous limiterons ici à la présentation de quelques moyens de valoriser l'information issue du croisement des classements individuels des sujets avec les classes de la partition centrale pour contribuer à l'examen de telles questions. Notons cependant que les aspects de positionnement des classes et de typicalité sont plus directement traités dans d'autres méthodes, comme les représentations par arbre à distance additive.

#### Dissimilarités entre classes

On dispose d'un indice de dissimilarité élémentaire entre objets pris deux à deux, qui est la proportion de sujets ayant classé les deux objets considérés dans deux classes différentes, soit :

$$dis(i, j) = [p - c_1(i, j)] / p \quad (6)$$

où  $p$  est le nombre de sujets (juges), et  $c_1(i, j)$  le nombre de sujets ayant classés  $i$  et  $j$  dans une même classe. On montre aisément que  $dis$  est un indice positif, symétrique, vérifiant  $dis(i, i) = 0$  pour tout  $i$ , et qu'il s'agit donc bien d'un indice de dissimilarité. Il est normalisé à 1 (et prend la valeur 0 pour deux objets toujours classés ensemble par les sujets, et la valeur 1 pour deux objets séparés dans deux classes par tous les sujets). On peut montrer également qu'il vérifie l'inégalité triangulaire ou propriété métrique.

Un pseudo-indice de dissimilarité entre deux classes peut être construit sur cette base : il s'agit de la proportion moyenne de sujets ayant classé séparément deux objets dont l'un appartient à la première classe considérée et l'autre à la seconde, en moyenne sur l'ensemble des couples d'objets ainsi constitués, soit :

$$Dis(CLA, CLB) = (1 / n_A) (1 / n_B) \sum_{i \in CLA} \sum_{j \in CLB} dis(i, j) \quad (7)$$

où  $CLA$  est une classe, d'effectif  $n_A$ , et  $CLB$  une autre classe, d'effectif  $n_B$ . Cet indice est positif, symétrique, normalisé à 1, et vérifie la propriété métrique, mais n'est pas un indice de dissimilarité car  $Dis(CLA, CLA)$  n'est pas nul. En pratique, dans les applications, on observe que  $Dis(CLA, CLA)$  est toujours nettement inférieur à  $Dis(CLB, CLC)$ , pour toute classe  $CLA$  et tout couple de classes  $(CLB, CLC)$ .

## Hétérogénéité des classes

On peut également définir un indicateur de l'hétérogénéité d'une classe de la partition centrale : la proportion moyenne de sujets ayant classé séparément deux objets distincts  $i$  et  $j$  appartenant tous deux à cette classe centrale, en moyenne sur l'ensemble des paires d'objets ainsi constituées :

$$H(CLA) = [1 / (n_A (n_A - 1) / 2)] \sum_{\substack{\{i,j\} \text{ paires} \\ \text{d'éléments de } CLA \\ \text{tels que } i \neq j}} dis(i, j) \quad (8)$$

Conventionnellement, pour une classe constituée d'un seul objet, on peut fixer  $H(CLA) = 0$ .

Cet indicateur peut théoriquement varier entre 0 et 1. En pratique, la partition centrale étant obtenue par optimisation à partir des partitions initiales des sujets, ses classes sont relativement homogènes et  $H$  prend des valeurs plus proches de 0 que de 1.

## Recherche d'exemplaires typiques dans les différentes classes

Il s'agit ici de trouver, parmi les objets d'une classe, ceux qui sont plus « typiques » que les autres, qui présentent une plus grande proximité avec la représentation catégorielle, avec le prototype supposé. L'analyse de la typicalité peut utiliser différentes voies, comme par exemple l'étude de descriptions verbales des exemplaires et catégories et de leur contenu sémantique (Dubois et Mazet, 1988, Collins et Loftus, 1975, Rosch et Mervis, 1975). Nous nous limitons ici aux informations que peuvent apporter les comparaisons par paires des objets (« classés ensemble » ou non) dans les partitions des différents sujets, concernant la typicalité de chaque objet relativement à sa catégorie.

A la différence des méthodes de représentation arborée, qui font émerger des nœuds latents qui par hypothèse rendent compte des prototypes, et permettent donc d'apprécier la proximité entre les exemplaires et l'hypothèse de prototype, les méthodes de partition centrale ne permettent pas de situer les prototypes par rapport aux exemplaires. Mais si la typicalité peut être appréciée par la proximité de l'exemplaire et du prototype supposé (ou par le nombre de traits communs entre exemplaire et catégorie, si l'on utilise les descriptions verbales), d'autres indices sont également utilisés, par exemple dans l'analyse des descriptions verbales, comme l'homogénéité inter-individuelle des propriétés évoquées pour chaque exemplaire : un exemplaire auxquelles de nombreuses propriétés sont attribuées, chacune par une majorité de sujets, est considéré comme plus typique (Dubois et Mazet, 1988, Dalhgren, 1985).

Sur la base d'un principe analogue d'homogénéité, si l'on ne considère que les données de classement et non les données verbales, on peut faire l'hypothèse qu'une bonne homogénéité inter-individuelle dans le rapprochement d'un exemplaire avec les exemplaires de la même classe centrale, tel qu'il ressort des

partitions des sujets, caractérise un exemplaire plus typique. Autrement dit, le fait qu'un objet soit plus fréquemment rapproché, par les différents sujets, des autres objets de la même classe de la partition centrale, peut être considéré comme un indice de typicalité positive. D'autre part, le fait qu'un objet ne soit que rarement rapproché d'autres objets n'appartenant pas à la classe constitue également un indice de typicalité, mais il s'agit alors plutôt d'une typicalité négative ou de contraste. La combinaison de ces deux informations peut être également utilisée, si l'on cherche une caractérisation plus globale de la typicalité.

Un indicateur possible de typicalité positive d'un objet est la proportion moyenne de sujets ayant classé cet objet avec un autre objet de la classe, en moyenne sur l'ensemble de ces autres objets :

$$Typpos(i) = [1 / (n_A - 1)] (1 / p) \sum_{\substack{j \in CLA \\ j \neq i}} c_1(i, j) \quad (9)$$

où  $i$  est un objet de la classe  $CLA$ . Il varie théoriquement entre 0 et 1, mais il est généralement plus proche de 1 que de 0, compte tenu de l'homogénéité relative des classes centrales.

De la même façon, on peut définir un indicateur de typicalité négative par la proportion moyenne de sujets n'ayant pas classé l'objet considéré avec d'autres objets n'appartenant pas à la même classe, en moyenne sur ces autres objets :

$$Typcontr(i) = [1 / (n - n_A)] (1 / p) \sum_{j \notin CLA} [1 - c_1(i, j)] \quad (10)$$

Cet indicateur varie théoriquement entre 0 et 1, mais il est généralement plus proche de 1.

Différents indicateurs combinant les deux indicateurs précédents peuvent être utilisés pour caractériser plus globalement la typicalité d'un objet. Par exemple, on peut utiliser la moyenne suivante :

$$Typglob(i) = [Typpos(i) + Typcontr(i)] / 2 \quad (11)$$

Cet indicateur varie théoriquement entre 0 et 1. Mais en fait, compte tenu de la façon dont les classes centrales ont été formées, il est généralement plus proche de 1. En pratique, il semble cependant prudent d'examiner également séparément les indicateurs de typicalité positive et de typicalité de contraste. D'autre part, l'analyse et l'interprétation des données verbales concernant les exemplaires et les catégories restent nécessaires pour identifier plus sûrement les exemplaires typiques.



## 4. Comparaisons entre partitions des différents sujets

Les résultats des études expérimentales évoquées dans l'introduction, reposant sur des partitions d'ensembles d'objets effectués par un échantillon de sujets, s'ils permettent de mettre en évidence des représentations catégorielles générales, donnent également des informations sur les différences inter-individuelles concernant les catégories. On peut aussi s'intéresser aux différences entre groupes de sujets aux caractéristiques différentes, ou correspondant à des conditions expérimentales différentes (voir par exemple Dubois *et al.*, 1993). Nous n'aborderons pas ici les moyens d'étudier ces différences par l'analyse des données verbales recueillies, mais nous nous limiterons à l'examen des informations que peuvent donner les partitions elles-mêmes concernant ces différences inter-sujets ou inter-groupes.

### Comparaison entre les partitions de deux sujets

Un indice de dissimilarité entre deux partitions (réalisées par deux sujets, par exemple) est le cardinal de la différence symétrique de leurs graphes, comme nous l'avons vu plus haut. Il a l'avantage d'être en cohérence avec les méthodes de partition centrale utilisées pour mettre en évidence les catégories générales, puisque ces méthodes utilisent la même métrique. De plus cette distance traduit directement le nombre de désaccords entre deux partitions.

Si l'on cherche plutôt à apprécier la proximité entre partitions, on peut aussi recourir à différents critères contingentiels connus, utilisés habituellement comme mesure d'association entre variables qualitatives, tels que les critères de Rand, du  $\chi^2$  (de l'écart à l'indépendance), de Belson, de l'écart à l'indépendance, de l'écart à l'indétermination, *etc.* (voir notamment Abdallah et Saporta, 1998). Parmi ceux-ci, le critère de Rand, que l'on peut écrire

$$R(P_i, P_j) = [2 \sum_{u \in P_i} \sum_{v \in P_j} (n_{u \cap v})^2 - \sum_{u \in P_i} n_u^2 - \sum_{v \in P_j} n_v^2 + n^2] / n^2 \quad (12)$$

où  $P_i$  et  $P_j$  représentent deux partitions du même ensemble de  $n$  objets, et  $n_u$ ,  $n_v$  et  $n_{u \cap v}$  les effectifs des classes  $u$ ,  $v$  et de  $u \cap v$ , respectivement, est un des seuls (avec le  $\chi^2$  normé par le coefficient de Tchuprow) à vérifier la condition :

$$R(P_i, P_i) = R(P_j, P_j) = R_{max} > R(P_i, P_j) \quad \text{pour toute paire d'éléments } P_i, P_j \quad i \neq j \\ \text{de l'ensemble des partitions}$$

Cet indice de similarité est normalisé à 1 car  $R_{max} = 1$ . Une valeur seuil de 1/2 peut être considérée comme caractéristique de la liaison entre deux variables (ou partitions) : si le critère de Rand est supérieur à cette valeur, les deux variables sont considérées comme liées, si l'on suit les propositions d'Abdallah et Saporta (1998). Cet indice a d'autre part l'intérêt d'être en bonne cohérence avec la distance du cardinal de la différence symétrique des graphes, utilisée pour

l'obtention de la partition centrale. En effet, on peut montrer que la distance du cardinal de la différence symétrique des graphes s'écrit également :

$$D(P_i, P_j) = \sum_{u \in P_i} n_u^2 + \sum_{v \in P_j} n_v^2 - 2 \sum_{u \in P_i} \sum_{v \in P_j} (n_{u \cap v})^2 \quad (13)$$

On montre d'autre part sans difficultés que la valeur maximale possible de cette distance est  $D_{max} = (n^2 - n)$  et qu'elle est atteinte pour la distance entre la partition la plus fine ( $n$  classes) et la partition la plus grossière (une seule classe). Si l'on déduit directement de cette distance un indice de similarité  $S(P_i, P_j)$ , normalisé à 1,

$$\begin{aligned} S(P_i, P_j) &= [D_{max} - D(P_i, P_j)] / D_{max} \quad (14) \\ &= [2 \sum_{u \in P_i} \sum_{v \in P_j} (n_{u \cap v})^2 - \sum_{u \in P_i} n_u^2 - \sum_{v \in P_j} n_v^2 + n^2 - n] / [n^2 - n] \end{aligned}$$

on peut voir qu'il est très semblable dans sa formulation au critère de Rand, et qu'il prend des valeurs très proches dès que  $n$  est faible devant  $n^2$ .

### Comparaisons entre groupes de sujets

Soient  $G1$  et  $G2$  deux sous-ensembles de l'ensemble des sujets, pouvant correspondre par exemple à des individus de caractéristiques différentes (expérimentés *versus* inexpérimentés, par exemple), ou à deux conditions expérimentales différentes.

La comparaison peut porter bien sûr sur les partitions centrales obtenues pour chacun des deux groupes. Les différences entre ces deux partitions peuvent être interprétées qualitativement, et la distance/proximité entre les deux peut être appréciée au moyen des critères évoqués plus haut, distance du cardinal de la différence symétrique des graphes ou critère de Rand, par exemple. Cette distance entre les partitions centrales correspondant à  $G1$  et  $G2$  peut aussi être comparée à la distance moyenne entre chacune de ces partitions centrales et les partitions dont elles sont tirées, ou aux distances entre partitions des sujets pris deux à deux au sein de l'ensemble  $G1$ , ou de l'ensemble  $G2$  (distances inter-individuelles).

Il est aussi possible, sans utiliser les partitions centrales, de comparer les distances entre partitions du groupe  $G1$  et partitions du groupe  $G2$  (plus précisément les distances au sein de couples de partitions  $P_{1i}, P_{2j}$  tels que  $P_{1i}$  est la partition effectuée par un sujet  $i$  du groupe  $G1$  et  $P_{2j}$  une partition effectuée par un sujet  $j$  du groupe  $G2$ ) aux distances intra-condition, c'est-à-dire au sein de couples de  $G1$  et au sein de couples de  $G2$ . La comparaison peut aussi porter sur des indices de similarité. La question peut être posée de savoir si les différences entre conditions ne sont pas susceptibles de résulter simplement d'effets d'échantillonnage, c'est-à-dire de différences entre les échantillons de sujets des différentes conditions (même si l'affectation des sujets aux différents groupes est effectuée

aléatoirement). Il est possible d'examiner cette question en s'appuyant sur des tests d'hypothèses (voir Brenac, 2002).

## **5. Exemple d'application**

Nous exposons ci-dessous quelques résultats d'une recherche portant sur la mise en évidence de représentations catégorielles de voies urbaines dans le contexte de la conduite automobile, selon différentes conditions (jour, soirée, nuit). Il ne s'agit que d'une présentation très partielle, destinée à illustrer les éléments de méthode exposés dans les points précédents. Cette recherche, conduite par M.C. Montel, fera en elle-même l'objet d'autres publications, qui détailleront la problématique générale de ces travaux, la nature des données et des dispositifs expérimentaux, et rendront compte complètement de l'interprétation des résultats, intégrant l'exploitation des données verbales recueillies à l'issue des expérimentations.

### **Brève présentation de l'expérimentation**

La recherche s'intéresse plus particulièrement aux catégories mentales de sites de voirie urbaine en conditions nocturnes, dans le but de pouvoir étudier ultérieurement les liens entre aménagements, catégories et attentes associées développées par les conducteurs, détection et identification (d'autres usagers peu visibles, notamment), compte tenu des incidences de ces éléments dans la production d'accidents de la circulation en période de nuit (Montel, 2000).

Un échantillon tiré au hasard de 65 sites de voirie urbaine sur le territoire d'une ville d'environ 250 000 habitants a fait l'objet de photographies à trois périodes de la journée : de jour, en soirée, et en milieu de nuit, en évitant que les photographies ne comportent des éléments trop spécifiques (présence d'un usager susceptible d'interagir de façon immédiate avec le conducteur, dont la photo rend le point de vue) afin qu'elles rendent plutôt compte des caractéristiques générales des voies de leur environnement et des activités qui s'y déroulent. Donc trois séries de photographies (65 en période diurne, 65 en soirée, 65 en pleine nuit) ont été obtenues. Chaque série a été présentée à 19 sujets ayant une certaine expérience de la conduite (permis de conduire datant de plus de cinq ans et pratique de la conduite). La consigne transmise à chaque sujet était de s'imaginer en situation de conducteur et de proposer un regroupement des sites jugés similaires compte tenu des situations ou problèmes qu'il pourrait être amené à gérer en conduisant sur de tels sites. Aucune contrainte sur le nombre de classes ou groupes n'était imposée. 57 sujets (19 par condition) ont donc produit chacun une partition de l'ensemble des 65 sites, un entretien étant ensuite réalisé avec chacun d'entre eux pour leur permettre d'explicitier ou de commenter les classes constituées.

## Comparaisons entre les partitions initiales

Nous ne présentons ici que les résultats concernant la période « Soirée » des résultats similaires ayant été obtenus pour la période « Jour » et pour la période « Nuit ».

Les nombres de classes obtenus par les 19 sujets de ce groupe sont les suivants : 9, 10, 5, 6, 14, 9, 11, 8, 9, 3, 5, 10, 12, 14, 5, 7, 15, 8, 7. (Soit par ordre croissant : 3, 5, 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 10, 10, 11, 12, 14, 14, 15). Cela reflète une certaine diversité, autour d'une moyenne qui est de l'ordre de 9. Les catégories correspondent *a priori* au niveau de base, et des différences dans le nombre de catégories au niveau de base peuvent correspondre à d'éventuelles différences inter-individuelles (entre sujets expérimentés et inexpérimentés, par exemple). Il n'est pas exclu cependant que certains sujets se soient placés, en dépit des précautions prises concernant la consigne, à un niveau de catégorisation superordonné ou sous-ordonné, introduisant un « bruit » susceptible de perturber les résultats.

Le tableau 1 représente une matrice dont chaque terme  $(s,t)$  représente le critère contingenciel de Rand rendant compte de la similarité entre la partition du sujet  $s$  et la partition du sujet  $t$ , pour les 19 sujets de la condition « Soirée ». On peut noter qu'il y a en général une bonne proximité entre les partitions, à la différence d'autres types d'application, en classification notamment, où les partitions ne rendent pas compte d'une répartition en classes effectuée dans des conditions identiques par un ensemble de sujets, mais du classement des objets d'étude en fonction des modalités prises par différentes variables qualitatives, de nature éventuellement très diverse.

Tableau 1. Matrice des critères de Rand entre partitions des sujets en condition « Soirée ». Chaque terme  $(s,t)$  de cette matrice représente le critère de Rand rendant compte de la similarité entre la partition d'un sujet  $s$  (en ligne) et celle d'un sujet  $t$  (en colonne).

1.00	0.78	0.71	0.79	0.85	0.76	0.85	0.78	0.80	0.64	0.76	0.82	0.81	0.82	0.73	0.79	0.81	0.81	0.82
0.78	1.00	0.76	0.78	0.81	0.73	0.80	0.77	0.78	0.71	0.73	0.80	0.80	0.79	0.73	0.77	0.81	0.79	0.80
0.71	0.76	1.00	0.72	0.75	0.70	0.76	0.75	0.71	0.73	0.68	0.74	0.72	0.73	0.71	0.70	0.73	0.73	0.73
0.79	0.78	0.72	1.00	0.85	0.76	0.85	0.80	0.89	0.74	0.82	0.83	0.82	0.88	0.77	0.81	0.82	0.79	0.87
0.85	0.81	0.75	0.85	1.00	0.81	0.90	0.83	0.85	0.69	0.80	0.88	0.88	0.89	0.79	0.83	0.88	0.85	0.87
0.76	0.73	0.70	0.76	0.81	1.00	0.77	0.84	0.73	0.69	0.76	0.81	0.75	0.80	0.74	0.76	0.78	0.77	0.77
0.85	0.80	0.76	0.85	0.90	0.77	1.00	0.82	0.89	0.70	0.80	0.86	0.88	0.86	0.81	0.82	0.87	0.84	0.87
0.78	0.77	0.75	0.80	0.83	0.84	0.82	1.00	0.80	0.72	0.77	0.83	0.78	0.82	0.79	0.78	0.80	0.78	0.80
0.80	0.78	0.71	0.89	0.85	0.73	0.89	0.80	1.00	0.71	0.81	0.83	0.83	0.85	0.77	0.81	0.83	0.80	0.86
0.64	0.71	0.73	0.74	0.69	0.69	0.70	0.72	0.71	1.00	0.69	0.69	0.68	0.71	0.69	0.68	0.69	0.69	0.70
0.76	0.73	0.68	0.82	0.80	0.76	0.80	0.77	0.81	0.69	1.00	0.80	0.78	0.80	0.73	0.77	0.80	0.77	0.79
0.82	0.80	0.74	0.83	0.88	0.81	0.86	0.83	0.83	0.69	0.80	1.00	0.85	0.85	0.76	0.81	0.86	0.84	0.85
0.81	0.80	0.72	0.82	0.88	0.75	0.88	0.78	0.83	0.68	0.78	0.85	1.00	0.86	0.75	0.81	0.86	0.82	0.83
0.82	0.79	0.73	0.88	0.89	0.80	0.86	0.82	0.85	0.71	0.80	0.85	0.86	1.00	0.77	0.83	0.85	0.81	0.86
0.73	0.73	0.71	0.77	0.79	0.74	0.81	0.79	0.77	0.69	0.73	0.76	0.75	0.77	1.00	0.75	0.78	0.75	0.77
0.79	0.77	0.70	0.81	0.83	0.76	0.82	0.78	0.81	0.68	0.77	0.81	0.81	0.83	0.75	1.00	0.82	0.79	0.78
0.81	0.81	0.73	0.82	0.88	0.78	0.87	0.80	0.83	0.69	0.80	0.86	0.86	0.85	0.78	0.82	1.00	0.86	0.83
0.81	0.79	0.73	0.79	0.85	0.77	0.84	0.78	0.80	0.69	0.77	0.84	0.82	0.81	0.75	0.79	0.86	1.00	0.80
0.82	0.80	0.73	0.87	0.87	0.77	0.87	0.80	0.86	0.70	0.79	0.85	0.83	0.86	0.77	0.78	0.83	0.80	1.00

### **Partition centrale pour la condition « Soirée »**

La partition centrale obtenue par optimisation comporte 10 classes à deux objets ou plus et 15 classes réduites à un singleton (que l'on peut aussi considérer comme des objets non classés). Les résultats sont donnés ci-dessous. Les numéros sont les numéros des sites et des photographies correspondantes. Les classes à deux objets ou plus sont les suivantes :

- $CL1 = \{1, 2, 3, 16\}$
- $CL2 = \{4, 6\}$
- $CL3 = \{5, 10\}$
- $CL4 = \{8, 18, 41, 62\}$
- $CL5 = \{9, 11, 12, 13, 14, 15\}$
- $CL6 = \{20, 22, 27, 29\}$
- $CL7 = \{21, 23, 24, 25, 26, 28, 30, 31, 35, 37, 55, 63\}$
- $CL8 = \{36, 39, 40, 46, 52, 59\}$
- $CL9 = \{42, 49, 50, 54, 56, 57, 64, 65\}$
- $CL10 = \{48, 51\}$ .

Les sites isolés sont les suivants : 7, 17, 19, 32, 33, 34, 38, 43, 44, 45, 47, 53, 58, 60, 61.

La proximité des numéros dans certaines classes tient à l'ordre des prises de vues photographiques (par exemple les photos 20 à 33 ont été prises dans le centre ancien de la ville). On voit que les catégories des sujets, à qui les photographies étaient données sous forme d'une pile classée dans un ordre aléatoire (différent pour chaque sujet), ne sont pas sans lien avec certaines cohérences spatiales, ou certaines catégories des expérimentateurs.

Si l'on considère que les objets isolés constituent des classes, le nombre de classes (qui est alors de 25) peut paraître élevé par rapport au nombre moyen de classes des sujets (9). Le nombre de classes n'étant pas imposé dans cette méthode, s'il est nettement plus élevé que le nombre moyen de classes des sujets, cela pourrait signifier que la partition centrale représente plus une solution de consensus que les représentations d'un hypothétique sujet moyen ou sujet type. On pourrait aussi évoquer la possibilité que le nombre de classes obtenues par un sujet résulte davantage des limites de l'empan mnémorique (contenance de la mémoire de travail) et des conditions de l'expérimentation que du nombre effectif de ses catégories du niveau de base en mémoire de long terme, qui pourrait être en fait bien plus élevé. La partition centrale pourrait alors constituer au contraire une meilleure représentation des catégories générales et de leur nombre.

Si l'on considère plutôt que les objets isolés ne constituent pas des classes, mais des exemplaires peu typiques situés aux frontières des classes, le nombre de 10 classes (non réduites à des singletons) paraît plus proche du nombre moyen de classes des sujets (9).

En fait, l'interprétation des classes conduirait plutôt à une position intermédiaire. En effet quelques sites isolés constituent sans doute des catégories peu représentées : le site 32, par exemple est identifié par les sujets comme une rue piétonne (en fait c'est une ruelle commerçante du centre ancien, mixte – c'est-à-dire ouverte à tous types d'utilisateurs mais sans trottoirs, sans ségrégation entre types d'utilisateurs) et aucun autre site ne correspond à une rue piétonne. On peut noter que, de jour, le site 32 n'est plus identifié comme rue piétonne (sans doute parce que la lumière du jour donne une vision plus complète de sa morphologie et de son environnement) mais regroupé avec les autres rues commerçantes du centre. D'autres sites isolés, nettement plus nombreux, comme le 33 par exemple (voir le point suivant), restent sans doute seuls parce qu'ils sont difficiles à catégoriser, éloignés des prototypes. Enfin d'autres sites isolés, comme le 43, sont peut être restés non classés du fait d'un élément perturbateur, par exemple sur ce site la présence d'un fourgon stationné sur le trottoir et débordant sur la chaussée, bien que les expérimentateurs aient pris soin d'éviter au maximum de tels éléments de perturbation lors des prises de vue.

Nous n'entrerons pas ici dans le détail de l'interprétation des différentes classes, qui sera présenté dans des publications ultérieures, et qui s'appuiera, non seulement sur l'interprétation des scènes visuelles par les expérimentateurs, mais aussi sur l'analyse des productions verbales des sujets. Mentionnons simplement schématiquement, sur la base de notre interprétation personnelle des photographies de sites, les types de voiries et d'environnement urbain auxquels semblent renvoyer quelques grandes classes de la partition centrale : la classe  $CL5 = \{9, 11, \dots, 15\}$  correspond à de grandes voiries de périphérie, aux caractéristiques assez routières, la classe  $CL4 = \{8, 18, 41, 62\}$  correspond à des voies principales plus proches du centre et plus insérées dans le tissu urbain, la classe  $CL6 = \{20, 22, 27, 29\}$  correspond plutôt aux rues commerçantes du centre ville, la classe  $CL7 = \{21, 23, 24, \dots, 63\}$  rassemble des rues étroites de desserte, principalement dans le centre historique, la classe  $CL8 = \{36, 39, 40, 46, 52, 59\}$  correspond à des voiries de desserte assez larges, plutôt en périphérie d'agglomération dans un environnement pavillonnaire ou de bâti paraissant peu élevé, la classe  $CL9 = \{42, 49, 50, 54, 56, 57, 64, 65\}$  rassemble des voiries de desserte dans un environnement d'habitat collectif de moyenne ou grande hauteur.

### ***Comparaison avec les résultats de l'algorithme des transferts***

L'application de l'algorithme des transferts, initialisé successivement au moyen des partitions initiales des sujets  $S2, S3, S6, S11, S14, S16, S17$ , et de la partition la plus fine ( $n$  classes, ici  $n = 65$ ) a conduit à la partition centrale (optimale) qui vient d'être exposée, telle qu'elle a été obtenue par la méthode d'optimisation. L'application de cet algorithme, initialisé successivement au moyen des partitions initiales des sujets  $S1, S4, S5, S7, S8, S9, S10, S12, S13, S15, S18, S19$  et de la partition la plus grossière (une seule classe), a conduit à quatre autres partitions, qui ne sont pas des partitions centrales mais de simples optimums locaux.

## Dissimilarités entre classes

Nous illustrons d'abord ici, concernant le cas de la classe singleton {33} de la partition centrale exposée au point précédent, un usage possible du pseudo-indice de dissimilarité entre classes présenté dans la partie méthodologique. Le tableau ci-dessous présente le pseudo-indice de dissimilarité normalisé à 1 entre le singleton {33} et les classes à deux éléments ou plus.

Classe	Dissimilarité relativement au singleton {33}	Classe	Dissimilarité relativement au singleton {33}
$CL1 = \{1, 2, 3, 16\}$	1,00	$CL7 = \{21, 23, 24, 25, 26, 28,$	
$CL2 = \{4, 6\}$	1,00	30, 31, 35, 37, 55, 63}	0,62
$CL3 = \{5, 10\}$	1,00	$CL8 = \{36, 39, 40, 46, 52, 59\}$	0,75
$CL4 = \{8, 18, 41, 62\}$	0,97	$CL9 = \{42, 49, 50, 54, 56,$	
$CL5 = \{9, 11, 12, 13, 14, 15\}$	1,00	57, 64, 65}	0,66
$CL6 = \{20, 22, 27, 29\}$	0,78	$CL10 = \{48, 51\}$	0,66

Tableau 2. Dissimilarités entre la classe {33} et les classes à plus de deux éléments.

Le site 33 correspond à une rue de desserte bordée du côté gauche de bâti ancien (maisons contiguës de deux étages), et de l'autre côté de stationnement en épi et d'un mur (qui peut faire penser à l'arrière d'un immeuble, mais il s'agit de l'envers des remparts de la ville).

On remarque que la classe {33} est très dissemblable des cinq premières classes de la colonne de gauche, ce qui n'est pas surprenant si l'on note que ces classes rassemblent des voies artérielles ou principales. Elle est moins éloignée de la classe  $CL6 = \{20, 22, 27, 29\}$  qui représente les rues commerçantes du centre, et de la classe  $CL8 = \{36, 39, \dots, 59\}$ , qui concerne des voies de desserte en périphérie, plutôt en zone pavillonnaire ou de bâti peu élevé. Mais on remarque surtout la valeur relativement faible de la dissimilarité entre la classe {33} et la classe  $CL7 = \{21, 23, \dots, 63\}$  qui rassemble des rues étroites de desserte principalement dans le centre historique — cette proximité relative pouvant sans doute s'expliquer notamment par l'aspect du bâti sur le côté gauche du site 33 — et entre la classe {33} et la classe  $CL9 = \{42, \dots, 65\}$ , qui rassemble des voiries de desserte dans un environnement d'habitat collectif, et dont le site 33 se rapproche par certaines de ses caractéristiques (mur pouvant faire penser à l'arrière d'un immeuble, stationnement en épi, fréquemment rencontré également sur les voies de desserte larges dans des zones d'habitat collectif en périphérie). La dissimilarité entre la classe {33} et la classe  $CL10 = \{48, 51\}$ , qui rassemble deux sites de voirie de desserte en périphérie d'agglomération où apparaissent à la fois des immeubles assez hauts et du bâti peu dense et de faible hauteur, est

également assez faible, ce qui pourrait s'expliquer par certains traits communs (aspect d'arrière d'immeuble également sur le site 48, par exemple).

Il semble donc que le site 33 soit à la frontière de plusieurs classes, peu typique ou atypique car associant des traits relevant de catégories différentes et contrastées, et ne soit pas à considérer comme un exemplaire, unique dans l'échantillon, d'une autre catégorie peu représentée.

Un examen plus global des dissimilarités entre classes peut aussi être effectué. On peut alors noter, par exemple, si l'on considère les dissimilarités entre les classes à deux éléments ou plus, prises deux à deux (tableau 3), que les plus fortes dissimilarités sont celles qui opposent une classe de l'ensemble *CL1*, *CL2*, *CL3*, *CL4*, *CL5* à une classe de l'ensemble *CL7* à *CL10*. Cela n'est pas surprenant puisque les classes *CL1* à *CL5* concernent plutôt des voies principales, hors du centre ancien, et que les classes *CL7* à *CL10* concernent plutôt des voies de desserte, dans le centre ancien ou en périphérie. On peut relever également que les dissimilarités sont toujours assez fortes entre la classe *CL7*, qui regroupe principalement des voies de desserte étroites dans le centre ancien (ruelles), et les autres classes, la classe *CL6* lui étant cependant un peu moins dissemblable. A un moindre degré, la même remarque peut être faite concernant la classe *CL6*, qui regroupe plutôt des rues commerçantes du centre, vis-à-vis des autres classes. Par contre, entre les classes *CL1* à *CL5*, comme entre les classes *CL8* à *CL10*, les dissimilarités sont moins marquées.

	<i>CL1</i>	<i>CL2</i>	<i>CL3</i>	<i>CL4</i>	<i>CL5</i>	<i>CL6</i>	<i>CL7</i>	<i>CL8</i>	<i>CL9</i>	<i>CL10</i>
<i>CL1</i>										
<i>CL2</i>		0.54								
<i>CL3</i>			0.78							
<i>CL4</i>			0.61	0.85						
<i>CL5</i>				0.54	0.64					
<i>CL6</i>					0.75	0.85				
<i>CL7</i>						0.89	1.00			
<i>CL8</i>							0.79	0.95		
<i>CL9</i>								0.89	0.95	
<i>CL10</i>									0.69	0.75

Tableau 3. Pseudo-indice de dissimilarité entre classes, pour les classes à deux éléments ou plus.



## Hétérogénéité des classes

Le tableau 4 donne, pour les classes à deux éléments ou plus les valeurs de l'indice d'hétérogénéité qui a été présenté dans la partie méthodologique. On peut relever que, pour les classes concernant plutôt les voies principales hors du centre ancien (classes *CL1* à *CL5*), l'hétérogénéité est relativement importante à l'exception de celle de la classe *CL5*, la plus représentée, et qui rassemble plutôt des voies artérielles de périphérie, aux caractéristiques assez routières, et relativement isolées du tissu urbain environnant. L'hétérogénéité des classes *CL6* (rassemblant des rues commerçantes du centre), *CL7* (où l'on trouve plutôt des rues de desserte étroites dans le centre) et *CL9* (qui regroupe plutôt des voiries de desserte dans un environnement d'habitat collectif en périphérie) est relativement faible.

Classe	Indice d'hétérogénéité	Classe	Indice d'hétérogénéité
<i>CL1</i>	0,44	<i>CL6</i>	0,25
<i>CL2</i>	0,47	<i>CL7</i>	0,30
<i>CL3</i>	0,42	<i>CL8</i>	0,41
<i>CL4</i>	0,44	<i>CL9</i>	0,33
<i>CL5</i>	0,35	<i>CL10</i>	0,37

Tableau 4. Hétérogénéité des classes.

L'hypothèse suivante serait à examiner, pour l'interprétation de ces résultats : (i) les tissus urbains anciens, façonnés par l'histoire à l'ère pré-automobile, et imposant de fortes contraintes pour l'aménagement de la voirie, engendreraient des environnements de conduite relativement homogènes et reconnaissables ; (ii) les voiries conçues à l'ère automobile, soit en s'abstrayant de l'environnement urbain et en s'appuyant sur des normes de conception (cas de certaines artères de périphérie), soit dans le cadre d'une planification conjointe de l'urbanisme et de la voirie (cas de certaines voiries de desserte dans des zones d'habitat collectif), engendreraient, du fait notamment d'une certaine standardisation ou de l'application de modèles urbains, des environnements de conduite également assez homogènes et reconnaissables ; et (iii) d'autres contextes, plus marqués par l'intégration conflictuelle des contraintes liées à l'automobile et des conceptions techniques de la voirie dans des tissus urbains pré-existants, ou se développant de façon peu planifiée, engendreraient des environnements de conduite plus diversifiés.

## Typicalité des exemplaires au sein d'une classe

Nous ne présentons ici qu'un exemple, concernant la typicalité des exemplaires relevant de la classe *CL7*, qui regroupe principalement des ruelles du centre ancien, sans commerces. Le tableau 5 présente les valeurs prises par les

indicateurs de typicalité positive, de typicalité de contraste et de typicalité globale présentés au point 3, pour les différents sites de cette classe.

Site	Indicateur de typicalité positive	Indicateur de typicalité de contraste	Indicateur global de typicalité
21	0,74	0,91	0,82
23	0,70	0,91	0,80
24	0,76	0,92	0,84
25	0,69	0,91	0,80
26	0,59	0,90	0,74
28	0,72	0,91	0,82
30	0,77	0,92	0,85
31	0,71	0,92	0,81
35	0,70	0,92	0,81
37	0,68	0,92	0,80
55	0,65	0,91	0,78
63	0,64	0,92	0,78

Tableau 5. Indicateurs de typicalité pour les exemplaires de la classe CL7.

On peut noter que les valeurs de l'indicateur de typicalité de contraste (typicalité négative) sont toutes assez proches, et élevées, ces exemplaires n'étant tous que rarement classés par les sujets avec des exemplaires extérieurs à cette classe. Ce qui est en cohérence avec ce que nous avons vu plus haut concernant les dissimilarités entre classes, cette classe se distinguant très clairement de toutes les autres. Les valeurs de l'indicateur de typicalité positive sont plus dispersées : certains exemplaires, comme le 26, le 63 ou le 55, sont moins souvent classés avec les autres exemplaires de cette classe CL7. Le site 26 se distingue des autres sites de cette classe sans doute parce que le bâti ancien des ruelles ne se retrouve que sur le côté droit, la chaussée, un peu plus large, étant bordée à gauche par les remparts de la ville et une file de véhicules stationnés. Les sites 55 et 63 sont bien des rues de desserte étroites, cependant ils ne sont pas situés dans le centre ancien mais dans la proche périphérie, comme le montre la nature du bâti environnant. Cela pourrait expliquer qu'ils apparaissent moins typiques pour les sujets.

Si l'on cherche des exemplaires plus typiques de cette classe, par exemple pour représenter cette catégorie de voie dans une expérimentation ultérieure, on pourrait retenir, au vu de ce tableau, les sites 24 ou 30. L'analyse précise des sites et des données verbales pourraient cependant apporter d'autres informations utiles pour effectuer un tel choix.

## Partitions centrales pour les conditions « Jour » et « Nuit »

Nous ne mentionnons ici que quelques aspects des résultats concernant ces deux conditions expérimentales, de façon à illustrer notamment le cas de solutions multiples. L'exposé complet et détaillé des résultats et interprétations concernant ces deux conditions et les comparaisons entre conditions expérimentales fera l'objet, comme nous l'avons dit plus haut, d'une autre publication.

Dans la condition « Jour » 19 autres sujets ont regroupé en classes 65 photographies représentant les mêmes sites que ceux évoqués plus haut, mais pris en période diurne. La recherche de partition centrale par optimisation concernant cette condition « Jour » donne deux solutions, comportant 13 classes à deux éléments ou plus. Les deux solutions ne diffèrent que quant à l'affectation du site 62, regroupé dans un cas (solution I) avec la petite classe {4,5} et restant isolé dans l'autre solution (II). L'application de l'algorithme des transferts, initialisé successivement au moyen des partitions des sujets  $J1, J2, J5, J8, J11, J15, J16, J19$ , et de la partition la plus grossière (une seule classe), aboutit dans chaque cas à la solution I déjà obtenue par optimisation. L'application de cet algorithme initialisé par la partition du sujet  $J18$  aboutit à la solution II déjà obtenue par optimisation. Pour les initialisations reposant sur les partitions des sujets  $J3, J4, J6, J7, J9, J10, J12, J13, J14, J17$ , et sur la partition la plus fine ( $n$  classes), on obtient cinq autres partitions, qui ne sont pas centrales mais sont de simples optimums locaux.

Dans la condition « Nuit » (photographies prises en milieu de nuit), la recherche de partition centrale par optimisation donne également des solutions multiples, comportant toutes 12 classes à deux éléments ou plus. Neuf solutions sont obtenues, qui se différencient ici également de façon marginale, et résultent de la combinaison de trois variantes relatives au classement des sites 40, 44 et 59 :

- $a$  : {7, 47, 40}, {44, 59},
- $b$  : {7, 47}, {40}, {44, 59},
- ou  $c$  : {7, 47}, {44}, {40, 59},

et de trois autres variantes relatives au classement des sites 38, 48 et 51 :

- $d$  : {38, 48, 51},
- $e$  : {38}, {48, 51},
- ou  $f$  : {48}, {38, 51}.

L'application de l'algorithme des transferts, initialisé successivement au moyen des partitions des sujets  $N2, N3, N4, N5, N7, N8, N9, N10, N14, N15$  et de la partition la plus fine, conduit à retrouver six des neuf partitions centrales obtenues par optimisation. Pour les initialisations reposant sur les partitions des sujets  $N1, N6, N11, N12, N13, N16, N17, N18, N19$  et sur la partition la plus grossière, on obtient neuf autres partitions, qui ne sont pas des partitions centrales mais seulement des optimums locaux.

### **Comparaisons entre conditions expérimentales « Soirée » « Jour » et « Nuit »**

Nous nous limiterons ici aux aspects quantitatifs, sans entrer dans l'interprétation des différences entre ces trois conditions expérimentales. Sont d'abord exposés ci-dessous quelques résultats concernant les distances entre les différentes partitions obtenues. La distance considérée est celle du cardinal de la différence symétrique des graphes. Cette distance, entre deux partitions d'un ensemble de 65 objets, prend une valeur comprise entre 0 et  $(n^2 - n) = 4160$ . Nous présentons ensuite quelques éléments de discussion relatifs à ces résultats.

a) La distance entre les deux partitions centrales obtenues pour la période « Jour » vaut 4. La distance entre les neuf partitions centrales obtenues pour la période « Nuit » prises deux à deux, varie entre 4 et 12. Par comparaison avec les résultats exposés en b) et c), cela semble confirmer que les partitions centrales multiples obtenues pour une même condition (cas de la condition « Jour » ou de la condition « Nuit » ne diffèrent que marginalement.

b) La distance moyenne entre les partitions des sujets J1 à J19 (condition « Jour ») et l'une ou l'autre des deux partitions centrales obtenues pour la condition « Jour » vaut 694,6. La distance moyenne entre les sujets S1 à S19 (condition « Soirée ») et la partition centrale obtenue pour la condition « Soirée » vaut 597,7. La distance moyenne entre les sujets M1 à M19 (sujets « Nuit ») et l'une quelconque des neuf partitions centrales obtenues pour la condition « Nuit » vaut 756,8.

c) La distance entre la partition centrale « Jour » n°1 (solution I) et la partition centrale « Soirée » vaut 242. La distance entre la partition « Jour » n°1 et la partition centrale « Nuit » correspondant à la combinaison  $a \times d$  (voir plus haut) vaut 212. La distance entre la partition centrale « Soirée » et la partition centrale « Nuit » correspondant à la variante  $a \times d$  vaut 138. Si l'on se base sur d'autres variantes de partitions centrales pour les conditions « Jour » ou « Nuit », ces résultats ne sont pas notablement modifiés, du fait des faibles distances entre ces variantes.

#### ***Différences entre conditions concernant la variabilité inter-individuelle***

Les résultats du b) tendraient à montrer que les différences inter-individuelles sont moins importantes concernant les catégories de voiries urbaines en période « Soirée » par comparaison avec les différences inter-individuelles en condition « Jour » et surtout en condition « Nuit ». Cela tient-il aux caractéristiques des scènes visuelles, qui seraient par exemple plus informatives en soirée que de jour du fait des indices émanant de sources lumineuses (vitrines, enseignes, fenêtres éclairées, etc.), et plus informatives en soirée que de nuit, parce que les activités y sont plus présentes et plus aisément décelables ? Cette information plus complète se traduirait par une réduction des ambiguïtés susceptibles de favoriser des

différences de classement entre sujets. Il s'agirait alors de différences dans la lisibilité, la « reconnaissabilité » des sites entre les différentes périodes, plus que de différences en matière de variabilité inter-individuelle des catégories elles-mêmes. Il n'est pas à exclure cependant que ces résultats traduisent simplement les effets d'échantillonnage, c'est-à-dire les effets des différences entre les trois échantillons de sujets correspondant aux trois conditions expérimentales, bien que les sujets aient été répartis aléatoirement entre ces trois échantillons.

### ***Différences entre conditions concernant les catégorisations***

Les résultats présentés en c) semblent mettre en évidence l'existence d'une différence en matière de catégorisation de sites de voirie urbaine selon la période de la journée. Les catégories correspondant à la période « Nuit » et à la période « Soirée » apparaissent assez proches, du point de vue quantitatif exprimé ici par la distance entre partitions centrales, relativement à celles de la période « Jour », situées à plus forte distance. Cependant, toutes ces différences restent moins importantes que les différences inter-individuelles, ce qui doit inciter à la prudence : les différences entre conditions ne sont-elles pas liées aussi aux effets d'échantillonnage, au moins en partie ? Une démarche qualitative, reposant sur l'analyse comparée et l'interprétation des classes des trois partitions centrales obtenues et sur l'interprétation des données verbales recueillies est susceptible d'apporter d'autres éléments concernant les différences entre conditions expérimentales.

## **6. Synthèse et conclusion**

Cette recherche portait sur quelques aspects quantitatifs des méthodes visant à mettre en évidence des représentations catégorielles générales en mémoire permanente, dans le cadre d'expérimentations consistant à demander à des sujets de classer des ensembles d'objets (objets simples, scènes, événements, *etc.*, présentés sous forme de photographies, de textes). De telles recherches, notamment lorsqu'elles sont réalisées dans le cadre de l'étude de la conduite de systèmes, privilégient souvent aujourd'hui la mise en évidence d'un « niveau de base » des catégories, au sein de l'organisation hiérarchique des connaissances en mémoire, niveau préféré correspondant au meilleur rapport entre coût cognitif et contenu informatif des catégories dans un contexte d'activité défini.

Nous avons d'abord examiné différentes méthodes permettant le passage des partitions individuelles, obtenues à l'issue des expérimentations, à une classification ou partition rendant compte de représentations catégorielles plus générales. Cet examen a montré que les méthodes de partition centrale présentent certains avantages, en particulier du fait qu'elles permettent un accès direct à des classes d'objets correspondant *a priori* à un niveau de base des catégories (sous réserve que le dispositif expérimental, et notamment la consigne, aient bien été

conçus pour que les partitions individuelles reflètent le niveau de base des catégories des sujets dans le contexte spécifié). Elles apparaissent de ce point de vue complémentaires des méthodes de représentation par arbre à distance additive, plus largement utilisées jusqu'à présent dans les recherches sur les catégories, et qui présentent des avantages à d'autres points de vue (accès à l'organisation hiérarchique et représentation de la typicalité des exemplaires).

Nous avons ensuite fait une présentation détaillée d'une méthode de partition centrale par la règle de majorité sous contraintes de Condorcet, utilisant une procédure de résolution exacte par programmation linéaire en nombres entiers due à Marcotorchino et Michaud (1982), et élaboré des programmes permettant sa mise en œuvre dans un environnement actuel. Ces programmes ont été validés par l'application à des jeux de données déjà utilisés par d'autres auteurs. Une application à un problème de grande taille ( $n = 65$  objets) a montré que les capacités des micro-ordinateurs ne constituent plus un obstacle à la mise en œuvre de telles méthodes, malgré la taille imposante de la matrice des contraintes (de l'ordre de  $n^2 \times n^3$ ).

Par rapport à d'autres méthodes de partitionnement ou de partition centrale, l'intérêt de cette méthode est qu'elle permet d'accéder à un optimum exact, donc à une partition réellement centrale, et non à un simple optimum local. Cet optimum exact n'est cependant pas nécessairement unique. Concernant la non-unicité possible de la solution, les exemples que nous avons traités montrent cependant, sans qu'on puisse en tirer de conclusion de portée plus générale, que la méthode de résolution par programmation linéaire permet d'accéder à des solutions multiples, et que celles-ci sont (dans ces cas) toujours très proches, ne différant que sur le classement d'un petit nombre d'objets. Nous avons également vérifié dans ces différents exemples, que l'heuristique de l'algorithme des transferts ne mettait pas en évidence d'autres solutions optimales concurrentes de celles déjà trouvées par optimisation.

Un certain nombre de compléments méthodologiques ont été ensuite mis au point, pour faciliter l'examen de la partition centrale obtenue, notamment concernant les dissimilarités entre classes centrales, leur homogénéité, et la typicalité des exemplaires au sein de chacune d'entre elles. D'autres éléments méthodologiques ont été également proposés, permettant d'utiliser la notion de distance entre partitions, qui est au fondement des méthodes de partition centrale, pour analyser d'autres aspects des données issues des expérimentations : différences inter-individuelles, dispersion, différences entre conditions expérimentales.

L'utilisation de l'ensemble des éléments de méthode présentés a été illustrée par une application à une recherche concernant les représentations catégorielles des conducteurs en matière de voiries urbaines selon différents moments de la journée

(jour, soirée, pleine nuit). Cet exemple a montré le caractère opératoire de la démarche proposée.

**Remerciements** : Ce travail méthodologique n'aurait pu être réalisé sans le support des travaux de recherche de M.-C. Montel sur les catégories de voiries urbaines. L'auteur tient à la remercier, ainsi que P. Van Elslande, C. Nachtergaële et J.-E. Michel, qui ont également participé à ces travaux. Il est également reconnaissant à S. Després, D. Dubois et J. Poitevineau, dont les commentaires ont permis d'améliorer notablement cet article.

## Références

- Abdallah, H. et Saporta, G. (1998) Classification d'un ensemble de variables qualitatives, *Revue de statistique appliquée*, 46, 4, 5-26.
- Baillargeon, G. (1996) *Programmation linéaire appliquée*, Trois-Rivières (Québec) , Les éditions SMG, 434 p.
- Barthélémy, J. P., Guénoche, A. (1988) *Les arbres et la représentation des proximités*, Masson, Paris, 240 p.
- Barthélémy, J. P., Luong, N.X. (1986) Représentations arborées de mesures de dissimilarité, *Statistiques et analyses de données*, 11, 1, 20-41.
- Brenac, T. (2002) *Contribution des méthodes de partition centrale à la mise en évidence expérimentale de catégories cognitives*, Rapport INRETS, Outils et méthodes, n°11, 92 p.
- Brenac, T., Fleury, D. (1999) Le concept de scénario type d'accident de la circulation et ses applications, *Recherche Transports Sécurité*, 63, 63-76.
- Buneman, P. (1971) The recovery of trees from measures of dissimilarity, in F. R. Hodson, D. G. Kendall, P. Tautu (éds.), *Mathematics in archeological and historical sciences*, Edinburgh University Press, 387-395.
- Cantor, N., Smith, E. E., French, R., Mezzich, J. (1980) Psychiatric diagnosis as prototype categorization, *Journal of abnormal psychology*, 89, 181-193.
- Cavalli-Sforza, L. L., Edwards, A. W. F. (1967) Phylogenetic analysis models and estimation procedures, *American Journal of Human Genetics*, 19, 233-25.
- Celeux, G. (1984) *Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement*, Rapport de recherche, n° 301, INRIA.
- Chandon, J. L., Lemaire, J., Pouget, P. (1980) Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, *RAIRO*, 14, 2, 157-170.
- Collins, A.M., Loftus, E.F. (1975) A spreading activation theory of semantic processing, *Psychological review*, 82, 407-428.
- Cunningham, J. P. (1978) Free trees as representations of psychological distances, *Journal of mathematical psychology*, 17, 165-188.
- Dahlgren, K. (1985) The cognitive structure of social categories, *Cognitive science*, 9, 379-398.
- Després, S. (1998) Recueil et analyse de scénarios génériques d'accidents, constitution et modélisation de scénarios consensus, Actes de la conférence « Ingénierie et connaissance 98 », Paris, 231-241.
- Dubois, D. (1991) *Sémantique et cognition*, Paris , Éditions du CNRS.
- Dubois, D. , Fleury, D (1994) From classification to cognitive categorization : the example of the road lexicon, in E. Diday, Y. Lechevallier, M. Schader, P. Bertrand et P. Burtschy (éds.), *New approaches in classification and data analysis*, Berlin , Springer-Verlag, 25-35.
- Dubois, D., Fleury, D., Mazet, C. (1993) Représentations catégorielles : perception et/ou action ? in A. Weil-Fassina, P. Rabardel et D. Dubois (éds.), *Représentations pour l'action*, Toulouse , Octares, 79-93.
- Dubois, D., Mazet, C. (1988) *Représentations catégorielles d'espaces routiers*, Rapport de l'Ecole pratique des hautes études, Laboratoire de psychologie du travail, 37 p.
- Fleury, D., Brenac, T. (2001) Accident prototypical scenarios, a tool for road safety research and diagnostic studies, *Accident analysis and prevention*, 33, 267-276.
- Fleury, D., Dubois D., Fline C., Peytavin, J. F. (1991) *Catégorisation mentale et sécurité des réseaux*, Rapport INRETS, n° 146, Arcueil, INRETS. 98 p.
- Fleury, D., Dubois, D., Morvant, C. (1993) *Expertise et structuration cognitive d'espaces routiers*, Rapport INRETS, n° 166, Arcueil, INRETS. 63 p.
- Guénoche, A. (1986) Représentations arborées des classifications, *RAIRO, Rech. Opér.*, 20, 341-353.



- Klotz, L. C., Blanken, R. L. (1981) A practical method for calculating evolutionary trees from sequence data, *Journal of theoretical biology*, 91, 261-272.
- Leplat, J. (1985) *Erreur humaine, fiabilité humaine dans le travail*, Paris, Armand Colin.
- Marcotorchino, J.F., Michaud, P. (1979) *Optimisation en analyse ordinale des données*, Paris, Masson, 211 p.
- Marcotorchino, J.F., Michaud, P. (1982) Agrégation de similarités en classification automatique. *Revue de statistique appliquée*, 30 (2), 21-44.
- Mazet, C., Dubois, D. Fleury, D. (1987) Catégorisation et interprétation de scènes visuelles : le cas de l'environnement urbain et routier, *Psychologie française*, 32, 85-96.
- Michaud, P. (1982) Agrégation à la majorité (I), *Hommage à Condorcet*, Paris, centre IBM France, Étude n° 51.
- Montel, M. C. (2000) Les accidents de nuit, in *Exploitations thématiques des Etudes détaillées d'accidents*, Rapport INRETS/DSCR 99-70-013-1, 55-102.
- Montjardet, B. (1990) Sur diverses formes de la « Règle de Condorcet » d'agrégation des préférences, *Mathématiques et sciences humaines*, 11, 61-71.
- Petit, J. L. (1993) Généralisation de la méthode des partitions centrales, *Revue de statistique appliquée*, XLI (3), 49-72.
- Rasmussen, J. (1974) *The human data processor as a system component, bits and pieces of a model*, Report Risø-M-1722, Risø National Laboratory (Danemark), 51 p.
- Régnier, S. (1983) Sur quelques aspects mathématiques des problèmes de classification automatique, *Mathématiques et sciences humaines*, 82, 61-71.
- Rosch, E. (1973) On the internal structure of perceptual and semantic categories, in T.E. Moore (éd.), *Cognitive development and acquisition of language*, New York, Academic Press, 111-144.
- Rosch, E. (1978a) Human categorization, in N. Warren (éd.), *Advances in cross-cultural psychology*, Vol. 1, Londres, Academic Press.
- Rosch, E. (1978b) Principles of categorization, in E. Rosch, B.B. Lloyd (éd.), *Cognition and categorization*, Hillsdale (N. J.), Erlbaum.
- Rosch, E., C.B. Mervis (1975) Family resemblances : studies in the internal structure of categories, *Cognitive psychology*, 7, 573-605.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., Boyes-Braem, P. (1976) Basic objects in natural categories, *Cognitive psychology*, 8, 382-439.
- Roux, M. (1998) Une comparaison objective mais limitée des algorithmes de constructions hiérarchiques, Actes des sixièmes rencontres de la Société Francophone de Classification, SFC'98, Montpellier, 21-23 septembre.
- Saporta, G. (1990) *Probabilités, analyse des données et statistique*, Paris, Technip. 493 p.
- Sattath, S., Tversky, A. (1977) Additive similarity trees, *Psychometrika*, 42, 3, 319-345.
- Smolenskii, Y.A. (1969) A method for linear recording of graphs. *USSR Computer Math. Phys.*, 2, 396-397.
- Tversky, A. (1977) Features of similarity, *Psychological review*, 84, 327-352.
- Tversky, B., Hemenway, K. (1983) Categories of environmental scenes, *Cognitive psychology*, 15, 121-149.
- Van Elslande, P. (2001) *Dynamique des connaissances, catégorisation et attentes dans une conduite humaine située. L'exemple des « erreurs accidentelles » en conduite automobile*, Thèse de doctorat, Univ. Paris V.
- Van Elslande, P., Alberton, L. (1997) *Scénarios-types de production de « l'erreur humaine » dans l'accident de la route, problématique et analyse qualitative*, Rapport INRETS, n° 218. Arcueil, INRETS. 180 p.
- Xu, Z. H. (1989) *Catégorisation de l'infrastructure par l'utilisateur en fonction des approches de l'analyse des données*, INRIA, Univ. Paris IX (sous la direction de E. Diday, D. Fleury), 111 p.
- Zaretskii, K. (1965) Construction d'un arbre sur la base d'un ensemble de distance entre ses feuilles (en russe) *Upekki Math. Nauk.*, 20, 90-92.



## Liste des auteurs

*Thierry Brenac* : INRETS

*Sophie David* : Modyco (CNRS-Université Paris X)

*Claude Marcel Hladik* : Eco-Anthropologie (CNRS-Museum National d'histoire naturelle).

*Bruno Lecoutre* : ERIS (CNRS-Université de Rouen), Site Colbert, 76821 Mont-Saint-Aignan Cedex, ([bruno.lecoutre@univ-rouen.fr](mailto:bruno.lecoutre@univ-rouen.fr))

*Marie-Paule Lecoutre* : ERIS-Laboratoire Psy.co (CNRS-Université de Rouen), UFR Psychologie, Sociologie, Sciences de l'éducation, 76821 Mont-Saint-Aignan Cedex. ([marie-paule.lecoutre@univ-rouen.fr](mailto:marie-paule.lecoutre@univ-rouen.fr))

*Patrick Pasquet* : Dynamique de l'évolution humaine (CNRS), 44, rue de l'Amiral Mouchez, Paris XIV

*Jacques Poitevineau* : Lcpe (CNRS), 44, rue de l'Amiral Mouchez, Paris XIV, ([jacques.poitevineau@ivry.cnrs.fr](mailto:jacques.poitevineau@ivry.cnrs.fr))

*Catherine Rouby* : UMR 5020 (CNRS-Université Lyon 1)

*Katia Rovira* : Laboratoire Psy.co (CNRS-Université de Rouen), UFR Psychologie, Sociologie, Sciences de l'éducation, 76821 Mont-Saint-Aignan Cedex. ([katia.rovira@univ-rouen.fr](mailto:katia.rovira@univ-rouen.fr))

## Numéros précédents

N°1 : Linguistique et ergonomie

N°2 : Linguistique et ergonomie II (Numéro en collaboration avec les Cahiers de Linguistique Sociale)

N°3 : Expériences individuelles et savoirs collectifs

N°4 : Espaces sensoriels et formes lexicales

N°5 : « Normes catégorielles » et listes lexicales

## À paraître

N° 7 : Espaces sensoriels et formes lexicales II : Juin 2003