Revue de Synthèse, IV^{e} S., 1-2, 1990. Dossier *Sciences Cognitives : quelques aspects problématiques.*

PRÉSENTATION

1. — INTRODUCTION

Ce numéro de la *Revue de synthèse* comprend un dossier sur les sciences cognitives dont le but est de proposer quelques remarques sur certains points délicats actuellement en débat.

Le développement considérable des sciences cognitives depuis plusieurs années est en train de bouleverser en profondeur le paysage de la recherche. Des remaniements interdisciplinaires sans doute irréversibles sont en cours. Ils transforment de façon significative les interfaces entre les mathématiques pures, la physique théorique, l'informatique et l'intelligence artificielle, les neurosciences, la psychologie et les sciences humaines. Depuis la biologie moléculaire dans les sciences de la vie et le structuralisme dans les sciences humaines, on n'avait pas connu de « révolution » scientifique d'une telle ampleur. Même le bilan le plus mesuré forcerait à parler de progrès rapides, convergents et décisifs des disciplines concernées. Les conséquences institutionnelles en sont de plus en plus manifestes : mise en place de grands programmes de recherche nationaux et internationaux, création de Groupements de recherche du C.N.R.S. et de nouveaux D.E.A., projets d'Instituts, Écoles d'été, etc. Elles ne peuvent qu'influencer et modifier de façon déterminante et durable nombre d'orientations. En ce qui concerne, en particulier, les sciences humaines, on peut sans doute affirmer, sans grand risque d'être démenti par le futur, que celles qui refuseront le « tournant cognitif » prendront le risque de se trouver irrémédiablement archaïsées.

Il existe d'excellentes introductions, récemment parues, à la situation nationale, européenne et internationale des sciences cognitives. Citons, entre autres, le rapport de Michel Imbert « Cognitive Science in Europe » (Springer Verlag, 1987), le dossier « Une nouvelle science de l'esprit » dans *Le Débat* n° 47, le dossier « Un tournant cognitif dans les sciences humaines » dans *Préfaces* n° 10, ainsi que le rapport « Sciences

Revue de synthèse: IV[e] S. N[os] 1-2, janvier-juin 1990.

cognitives » du *Programme de recherche sur les sciences de la communication* du C.N.R.S. (octobre 1989). Le lecteur intéressé y trouvera les renseignements scientifiques, les références bibliographiques et les éléments institutionnels de base.

Ce dossier est d'un genre différent. Plutôt que de présenter un panorama ou d'essayer de faire le point sur certains résultats expérimentaux assurés des neurosciences et de la psychologie cognitive, il se propose de discuter relativement en détail quelques-uns des délicats problèmes, de nature théorique et/ou épistémologique, faisant actuellement l'objet de débats particulièrement vifs dans la communauté cognitiviste.

2. — QUELQUES CARACTÈRES GÉNÉRAUX DES SCIENCES COGNITIVES[1]

Il n'est pas étonnant que les sciences cognitives débouchent sur nombre de difficultés concernant leur conceptualisation théorique, leur formulation et leur modélisation mathématiques ainsi que leur compréhension épistémologique. En effet, elles approchent en termes de sciences *naturelles* les structures, les actes et les processus mentaux constitutifs de facultés et d'objets comme la représentation (problème de l'existence, du statut, de la fonction des représentations mentales), l'intentionalité (problème de la capacité qu'ont les représentations de référer à l'environnement), le raisonnement (développement de systèmes formels modélisant correctement les processus psychologiques du raisonnement naturel : logiques floues, non monotones, etc.), la nature des concepts (problème des rapports concept-percept, catégorisation, schèmes, prototypes), l'apprentissage (développement de théories formelles), la communication (conçue comme transmission verbale et non verbale de représentations), l'architecture fonctionnelle de l'esprit (systèmes périphériques modulaires et systèmes centraux non modulaires, hypothèse fonctionnaliste, modèles connexionnistes). Cette approche est plus particulièrement effective dans les quatre domaines du langage, de la vision, du raisonnement et de l'action. Les sciences cognitives y étudient soit le développement, soit l'état adulte constitué, et se situent aux niveaux de réalité et de modélisation biologique, psychologique, algorithmique et formelle.

---

1. Ce paragraphe reprend en partie la contribution — dont nous avons assumé la responsabilité avec notre collègue Claude Debru — de la Commission de philosophie au Groupe thématique « Sciences cognitives et Communication » du Rapport de conjoncture du C.N.R.S. (décembre 1989).

Les sciences cognitives étant par définition interdisciplinaires, elles favorisent des interactions privilégiées, au nombre desquelles se trouvent les interactions entre la philosophie, la logique, la linguistique et la psychologie. Le rapport entre sciences cognitives et épistémologie est également privilégié pour deux raisons. D'un côté, les sciences cognitives sont des sciences jeunes, en plein processus de constitution de leurs objets, de leurs principes et de leurs méthodes. Elles rencontrent donc naturellement tous les problèmes de philosophie des sciences (ontologiques, épistémologiques, méthodologiques) que rencontrent les sciences en voie de constitution ou de refondation. Mais, d'un autre côté, ces sciences représentent la tentative de comprendre comme des processus *naturels,* expérimentables et mathématisables, un ensemble de phénomènes sur lesquels l'épistémologie n'a cessé de se pencher.

Leur lien avec l'épistémologie n'est donc pas extrinsèque. Celle-ci n'y intervient pas seulement, comme dans les autres sciences naturelles, à un niveau second, réflexif et critique. Elle y intervient d'emblée, de façon constitutive, dès le niveau des objets. Les concepts que la philosophie des sciences avait dû créer et développer pour penser les sciences d'objets sont à leur tour devenus des objets de science. Les thèmes évoqués plus haut, ainsi que ceux par exemple de l'innéisme des structures mentales, du jugement, de l'induction, de la corrélation entre les actes mentaux et leurs objets, de l'inférence non démonstrative, des rapports entre syntaxe et sémantique, entre extensionalité et intensionalité, etc., sont des thèmes qui sont au cœur des traditions philosophiques les plus anciennes et les plus centrales, qui ont été repensés de nombreuses fois, du cartésianisme à la phénoménologie, dans des optiques différentes et qui, dans la philosophie des sciences et la philosophie de l'esprit (avec les apports, problématiques, de la philosophie de la logique et de la philosophie analytique modernes), ont été en partie modélisés et formalisés. Les sciences cognitives contemporaines visent essentiellement à apporter une théorie naturaliste des actes mentaux correspondants.

Le paradigme actuellement dominant en sciences cognitives est celui du mentalisme computationnel et propositionaliste (Jerry Fodor et Zénon Pylyshyn en sont des représentants typiques). Il admet les thèses du fonctionnalisme (possibilité de séparer les problèmes de structure formelle et de calcul des représentations mentales de ceux de l'implémentation dans un substrat matériel, en particulier neuronal), ainsi que celles du solipsisme méthodologique (les rapports entre l'organisme et l'environnement n'étant pas entièrement régis par des lois causales, il est impossible d'admettre dans les hypothèses d'une psychologie scientifique une référence constitutive aux structures de l'environnement). On fait l'hypothèse qu'il existe un « langage de la pensée » constitué de repré-

sentations mentales symboliques possédant la structure d'un langage formel (symboles, expressions structurées syntaxiquement, règles de déduction, etc.). Ces représentations (structurées à plusieurs niveaux différents reliés par compilation) permettent de traiter des informations externes et internes. A travers un tel traitement, le monde physique objectif devient un monde structuré qualitativement par la perception, le langage et l'action, le monde de l'expérience phénoménologique. Ce paradigme est désormais classique et il existe de très nombreux travaux qui en développent les thèses. Mais nombreuses sont également les critiques. Elles ont conduit au développement d'un certain nombre de thèmes émergents originaux. Nous en examinerons brièvement quelques-uns : le connexionnisme, l'intentionalité perceptive, l'ontologie qualitative dans ses rapports à « l'écologie » perceptive.

(a) Développé par Hilary Putnam et Jerry Fodor, le fonctionnalisme est la « solution » donnée au problème des rapports entre les états mentaux et les états cérébraux qui a recueilli l'approbation de la majorité des philosophes et des spécialistes en sciences cognitives d'orientation formaliste. Il repose sur l'observation du fait qu'un programme informatique, qui est un ensemble d'instructions logiques, peut être « implémenté » dans des ordinateurs ayant des structures physiques différentes. Autrement dit, il repose sur la double analogie entre, d'une part, les étapes logiques d'un programme informatique et les états mentaux, et d'autre part, les états physiques d'un ordinateur et les états cérébraux. Les fonctionnalistes soutiennent que le vocabulaire et les concepts destinés à décrire, expliquer et prédire le comportement des états répertoriés comme « états cérébraux » ne sont pas *ipso facto* appropriés pour décrire, expliquer et prédire le comportement des états répertoriés comme « états mentaux ». Cette position non réductionniste permet une division rationnelle entre neurosciences et sciences cognitives mais ne permet pas de rendre compte du caractère qualitatif de certains états mentaux qui se définissent moins par leur rôle computationnel ou inférentiel que par la qualité de l'expérience qui leur est concomitante. En outre, il n'est pas assuré que l'on puisse séparer dans les machines cognitives biologiques comme le cerveau les problèmes de logiciel (« software ») et de matériel (« hardware »).

Les modèles connexionnistes qui distribuent les représentations sur des réseaux de micro-unités et de micro-traits relèvent d'un paradigme différent du paradigme classique. L'analogie est ici plus thermodynamique qu'informatique. On suppose que les systèmes cognitifs possèdent deux niveaux : un niveau macroscopique symbolique où l'on décrit formellement les structures et les mécanismes de la compétence et un niveau

microscopique sous-symbolique où l'on explique les processus dynamiques de la performance en termes mathématiques de systèmes dynamiques sur des réseaux.

Le connexionnisme est une théorie relevant d'une épistémologie de l'émergence : on pose que les structures formelles de la compétence que décrit le cognitivisme symbolique sont des organisations et des régularités émergeant des processus dynamiques sous-jacents. Il rencontre toutes les difficultés philosophiques des théories antiréductionnistes de l'émergence. Une question fort débattue actuellement est de savoir si le connexionnisme fournit en fait seulement une théorie de l'implémentation des algorithmes symboliques classiques dans des machines massivement parallèles ou bien s'il fournit, à l'intérieur même de l'hypothèse fonctionnaliste, une alternative proprement cognitive au cognitivisme classique.

(b) Des travaux considérables ont été consacrés au problème de l'intentionalité des représentations. En général (Searle, Dreyfus, Føllesdal, Dennett, etc.), l'approche est sémantique. Mais le problème critique est celui de l'intentionalité *perceptive*. Des travaux récents, portant en particulier sur la perception visuelle, permettent de mieux l'aborder. L'analyse du pattern rétinien bidimensionnel (en particulier, la détection des discontinuités qualitatives) permet de reconstruire d'abord les contours des objets et leur remplissement par des qualités sensibles, et ensuite le caractère volumique des objets occupant l'espace extérieur tridimensionnel et s'y déplaçant. Comme l'ont montré H. Dreyfus et D. Føllesdal, ces analyses cognitives sont étonnamment proches des analyses de la phénoménologie husserlienne sur le noème de la perception. De façon générale, les thèmes néo-aristotéliciens essentiels de la phénoménologie et de la *Gestalttheorie* sont actuellement repris et reformulés en termes neurocognitifs.

(c) Concernant le type de structure que l'on doit admettre dans l'environnement, deux points de vue s'opposent. Pour les cognitivistes classiques, le contenu objectif des structures de l'environnement est physique en un sens physicaliste et réductionniste. La façon dont le monde de l'expérience phénoménale est qualitativement structuré en formes, qualités, choses, états de choses, événements, processus, etc., qui sont appréhendés dans la perception et linguistiquement descriptibles, est le résultat d'une construction cognitive, c'est-à-dire de la façon dont nous nous représentons l'information physique externe (ondes lumineuses, sonores, etc.). On admet bien une « ontologie qualitative » (d'une façon ou d'une autre néo-aristotélicienne), mais celle-ci est conçue comme l'ontologie

d'un « monde projeté », monde sans contenu physique qualitatif, morphologique et structural, monde purement corrélatif des actes (des calculs) du sujet cognitif. A l'opposé de ces points de vue classiques, physiquement réductionnistes et cognitivement constructivistes et projectivistes (et donc nominalistes), il existe des points de vue plus réalistes attribuant des structures qualitatives, morphologiques et structurales à l'environnement, structures objectives émergeant de la physique sous-jacente. Dans le domaine visuel, un exemple en est fourni par la théorie « écologique » de Gibson.

Comme l'ont bien montré J. Fodor et Z. Pylyshyn, Gibson n'arrive cependant pas à définir cette objectivité non strictement physique de l'ontologie qualitative. En fait, sous-jacent à ce débat, il y a un des plus anciens et des plus lourds problèmes de la philosophie, celui du rapport entre objectivité physique et ontologie qualitative. Comme l'a explicité André Robinet dans son ouvrage récent sur Leibniz, cette « architectonique disjonctive » entre un mécanisme physicaliste et un systémisme structural est central pour l'ensemble de la pensée leibnizienne. Elle l'est encore pour la phénoménologie et la *Gestaltheorie*. On peut dire que c'est elle qu'il s'agit de dialectiser. L'écologisme en est un aspect.

Ces dernières années les choses se sont notablement transformées. D'abord les physiciens et les mathématiciens ont montré qu'il existe effectivement dans l'environnement des structures qualitatives émergentes (phénomènes critiques, catastrophes, structures dissipatives, etc.). Ils en ont fourni de nombreux modèles morphodynamiques. D'autre part, des spécialistes de la perception visuelle comme David Marr ont montré que les points de vue « écologiques » étaient compatibles avec une théorie computationnelle des processus de traitement de l'information. Enfin, les travaux sur la physique « naïve » et surtout sur la physique « qualitative » ont retrouvé, à partir d'une analyse du coût computationnel de la résolution des problèmes physiques, les résultats déjà anciens des approches morphodynamiques.

On peut conclure de la convergence remarquable de ces diverses approches, qu'il est désormais possible de développer une authentique ontologie qualitative qui soit le résultat à la fois d'une auto-structuration qualitative et morphologique spontanée (émergente) de l'environnement et d'une construction cognitive. Cela permet de sortir du solipsisme méthodologique en réfutant l'idéalisme propre à tous les mentalismes de la représentation. Le gain philosophique et scientifique est évident et considérable. Cela permet aussi de redéployer des conceptions en partie réalistes de la perception et du langage : ceux-ci explicitent des structures qualitatives possédant un contenu en grande partie objectif. Des travaux de linguistes cognitivistes comme L. Talmy, R. Jackendoff et R. Lan-

gacker vont dans ce sens. Cela vient remettre en partie en cause, dans le cadre d'une orientation néo-phénoménologique, le nominalisme logiciste propre à la tradition sémantique.

3. — LE PROBLÈME CENTRAL DU LIEN ENTRE LE PHYSIQUE ET LE SYMBOLIQUE

On voit que pour arriver à se constituer en sciences *naturelles* de l'esprit, les sciences cognitives doivent unifier *trois types d'objectivité* :

(i) l'objectivité *physique,* constitutive du monde externe (physique des signaux visuels et acoustiques, etc.) ;

(ii) l'objectivité *neuronale-informationnelle,* constitutive de la structure et de la fonction des dynamiques cérébrales (comme y a souvent insisté Jean-Pierre Changeux, il est illégitime de séparer structure et fonction) ;

(iii) l'objectivité *logico-symbolique,* constitutive des structures syntaxiques et sémantiques des systèmes symboliques et des langages formels en général.

Or ces trois objectivités sont fortement *hétérogènes* entre elles. En particulier, il existe *un véritable dualisme entre le physique et le symbolique.* Comment donc accéder à une théorie naturaliste unifiée ? On voit immédiatement comment des hypothèses fondamentales concernant le statut des états et des processus cognitifs pourront découler naturellement de la perspective adoptée sur cette question.

Si l'on part par exemple de la thèse que la seule objectivité naturelle se réduit à l'objectivité physique (thèse éliminationniste), on refusera alors d'accorder une réalité autre que descriptive et artéfactuelle aux représentations et aux actes mentaux et l'on identifiera ceux-ci à des états et à des processus neuronaux. Si l'on admet au contraire que l'objectivité logico-symbolique est *en tant que telle* réalisable dans des systèmes physiques traitant de l'information (thèse de l'I.A.), on pensera les systèmes cognitifs sur la base d'une analogie avec les ordinateurs. Si l'on pense qu'il existe bien des représentations mentales mais que l'objectivité logico-symbolique n'existe pas dans la nature, on développera par exemple l'hypothèse centrale introduite par Ch. Zeeman et R. Thom il y a déjà plus de vingt ans, à savoir que les unités sémantiques sont descriptibles par des attracteurs de dynamiques neuronales et que les structures syntaxiques s'ancrent dans des bifurcations de ces attracteurs (c'est une thèse analogue que développent actuellement de facon technique les modèles connexionnistes de réseaux de neurones formels).

Le débat sur des thèses de ce genre conduit immédiatement à nombre de problèmes techniques dont certains sont abordés ici.

4. — L'ORGANISATION DU DOSSIER

4.1. Joëlle Proust engage sa réflexion à partir d'un des problèmes épistémologiques centraux évoqués plus haut. Si, dans une optique mécaniste et formaliste, on identifie les actes cognitifs à des processus de traitement de l'information réductibles à des manipulations réglées de symboles, comment rendre compte du *contenu* des symboles ? Si les règles et la causalité des systèmes cognitifs sont purement syntaxiques, comment comprendre l'intentionalité, c'est-à-dire le rapport à des contenus, la directionalité de la conscience vers des objets, bref les objectivités immanentes (ce que Husserl considérait comme « le problème des problèmes ») ? Après avoir rappelé deux grandes traditions, celle des diverses formes de solipsisme méthodologique et de rationalisme (Descartes, Berkeley, Hume, Kant, Carnap : l'identification des contenus mentaux est méthodologiquement indépendante de leurs causes extérieures) et celle du naturalisme (Dewey, Quine : les contenus mentaux ne sont que des réponses de l'organisme à des états de choses externes), Joëlle Proust expose les thèses de Jerry Fodor. Selon Fodor, tout ce qui est cognitivement significatif est de nature computationnelle et seules sont déterminantes pour le système cognitif les propriétés formelles des représentations symboliques. Les opérations et processus mentaux sont par conséquent opaques, fermés à la sémantique de leurs symboles formels, même si les représentations mentales possèdent des propriétés sémantiques. D'où le solipsisme méthodologique fonctionnaliste défendu, nous l'avons vu, par Fodor. Comme on ne peut pas expliquer *nomologiquement* les relations causales entre les représentations mentales et leurs référents, il est impossible de conférer un statut scientifique aux thèses naturalistes. Joëlle Proust analyse alors le débat entre Fodor et Putnam à ce sujet et la façon dont Fodor en arrive à la conclusion que le contenu *étroit* (cognitif) des représentations mentales *n'est pas spécifiable dans la langue naturelle,* celle-ci l'ancrant nécessairement dans un contexte. Ce contenu étroit se distingue du contenu *large* (informationnel) qui, lui, est causalement solidaire des états de choses externes.

4.2. Après avoir rappelé les bases de la conception symbolique du cognitif, Jean-Pierre Desclés s'attache à montrer dans quelle mesure

l'histoire, la théorie mathématique et l'ingénierie des langages informatiques de haut niveau permet de mieux comprendre quel peut être le statut des représentations mentales. L'idée directrice de sa réflexion est que, entre, d'un côté, le support neuronal où les actes et les processus mentaux se trouvent physiquement implémentés et, d'un autre côté, les représentations symboliques de haut niveau (systèmes logiques, algorithmes, règles, procédures, stratégies, représentation des connaissances, scènes, plans, etc.), il existe une hiérarchie de niveaux *intermédiaires* reliés entre eux par des rapports de *compilation.* Après avoir fait l'historique de la compilation et des langages informatiques (FORTRAN, ALGOL, PASCAL, LISP, PROLOG, etc.), Jean-Pierre Desclés explicite l'acquis fondamental que représente selon lui le « principe de compilation » dans la compréhension du rapport entre le physique et le symbolique.

4.3. Quant à Massimo Piattelli-Palmarini, il se propose d'élucider la façon dont les énoncés peuvent *causer* des actions et, en particulier, le rôle du lexique dans cette *causalité sémantique.* Selon lui, la causalité sémantique est *naturelle,* bien que non physique. Mais comment accéder à une compréhension naturaliste *non réductionniste* du fait que la sélection sémantique n'est pas explicable par une liste finie et figée de correspondances entre énoncés et significations ? L'idée directrice est celle d'un *innéisme sémantique.* Elle ouvre à une théorie évolutionniste, causale et naturaliste, à la fois antiformaliste et antibehavioriste. Conformément aux thèses de Chomsky et de Fodor, il existerait un ensemble d'« atomes » et de contraintes sémantiques cruciaux qui seraient innés (donc tacites et inaccessibles à l'introspection) et qui conditionneraient universellement la structure des langues naturelles humainement accessibles.

4.4. Daniel Andler consacre son intervention au débat, actuellement d'une grande vivacité, entre le cognitivisme symbolique classique et le connexionnisme. Après avoir rappelé les différences caractéristiques opposant ces deux paradigmes (et certains de leurs sous-paradigmes), il montre que chacun est adapté à la conceptualisation et à la modélisation mathématique de certains aspects fondamentaux des systèmes cognitifs humains. Il est donc légitime de faire l'hypothèse d'une hiérarchie de niveaux cognitifs conduisant de « bas » niveaux de nature dynamique, associationniste et connexionniste à des « hauts » niveaux de nature formelle, inférentielle et symbolique.

4.5. En prenant comme exemple celui de la perception des couleurs, Francisco Varela développe un point de vue « écologique » et dyna-

mique, se démarquant des points de vue symboliques, formalistes et computationnels classiques. Nombre d'arguments expérimentaux réfutent la possibilité d'une réduction objectiviste de la perception de couleurs au traitement d'informations physiques comme la composition spectrale de la lumière ou la réflectance spectrale des surfaces visibles. Les espaces perceptifs de couleurs ne sont pas objectifs. Ils sont relatifs et définis par des domaines chromatiques spécifiques des systèmes visuels des différentes espèces animales. Selon Francisco Varela le défaut de l'objectivisme est de négliger le fait que l'environnement d'un organisme n'est pas pré-donné mais évolutivement spécifié et conformé. D'où une théorie de *l'énaction*. Le monde visuel n'est ni trouvé, ni inventé mais énacté.

4.6. Enfin, également à propos de l'exemple de la perception visuelle, nous développons la thèse qu'il n'est pas possible de résoudre le problème central du rapport entre le physique et le symbolique sans introduire un niveau intermédiaire, de nature proprement *morphologique* et mathématisé au moyen de modèles morphodynamiques. Selon nous, le point de vue connexionniste ne saurait suffire pour constituer et élaborer mathématiquement les sciences cognitives en tant que sciences naturelles. Car encore faut-il comprendre comment des structures qualitatives *intrinsèquement* significatives peuvent émerger dynamiquement et stablement des substrats physiques, tant internes (neuronaux) qu'*externes*. En convergence avec les théories de David Marr et de Jan Koenderink, nous montrons comment certains modèles morphodynamiques de ces structures morphologiques émergentes peuvent être intégrés aux théories perceptives.

***

Je remercie la *Revue de synthèse* et Ernest Coumet de m'avoir proposé d'organiser ce dossier. J'exprime également ma gratitude aux auteurs des contributions. J'espère que la diversité de leurs points de vue permettra au lecteur d'apercevoir la richesse et la fécondité de l'interdiscipline cognitiviste depuis que les théories logiques, les modèles physiques et les mathématiques morphodynamiques s'y conjuguent constructivement avec les neurosciences et la psychologie. A n'en pas douter, l'enjeu est historique pour les sciences humaines.

Jean PETITOT.

DE LA DIFFICULTÉ D'ÊTRE NATURALISTE EN MATIÈRE D'INTENTIONALITÉ *

Ce qu'on appelle les « sciences cognitives » constitue aujourd'hui une nébuleuse de disciplines et de méthodologies différentes, dans laquelle toutefois émergent certaines hypothèses générales communes à la majorité des chercheurs. L'une d'entre elles consiste à postuler que connaître, c'est « traiter » de l'information, c'est-à-dire manipuler des symboles de manière réglée, conformément à des processus qui sont de nature *formelle.* Il est évident que l'un des problèmes fondamentaux que les sciences cognitives doivent résoudre est celui de savoir comment les symboles qui sont censés rendre possibles les processus mentaux peuvent recevoir un *contenu,* c'est-à-dire un sens. Ce problème se pose de façon particulièrement exemplaire dans le cas de l'Intelligence Artificielle : comment un système informatique, soit un ensemble de règles de type *syntaxique,* peut-il prétendre modéliser des processus où l'interprétation du monde joue un rôle essentiel — dépasser la simulation entendue comme bricolage ? De tels systèmes ne sont-ils pas voués à rester purement formels, essentiellement dépourvus de sens en dépit de la tendance de l'utilisateur à projeter du sens sur les formules produites ? Mais le problème se pose de façon tout aussi urgente dans le cas de la linguistique, de la théorie de la vision, ou de la théorie de l'action : si les processus cognitifs de l'esprit — élémentaires ou supérieurs — sont de nature mécanique (c'est-à-dire s'ils sont régis par des règles formelles s'attachant à la seule forme des symboles), *il faut bien* que ces symboles représentent quelque chose, se voient conférer un sens. C'est là le problème dit de l' « intentionalité ».

Notons immédiatement que lorsqu'on parle d' « intentionalité » dans ce contexte, on s'écarte du sens courant en vertu duquel est intentionnel un acte prémédité ou réfléchi, pour retrouver le sens philosophique

* Je remercie Pierre Jacob d'avoir relu cet article et de m'avoir suggéré des éclaircissements ou des précisions sur plusieurs points.

traditionnel, des scolastiques aux phénoménologues : Brentano commente le terme d' « intentionalité » en termes de

> « rapport à un contenu, direction vers un objet (sans qu'il faille entendre par là une réalité), ou objectivité immanente [....] Dans la représentation, c'est quelque chose qui est représenté, dans le jugement quelque chose qui est admis ou rejeté, dans l'amour quelque chose qui est aimé, dans la haine quelque chose qui est haï, dans le désir quelque chose qui est désiré, et ainsi de suite »[1].

L'intentionalité est ainsi le lien sémantique par excellence : dire qu'un symbole *représente*, dire qu'il *renvoie* à quelque chose et, de façon plus générale, dire qu'un état mental *porte sur* un contenu sont autant de façons d'exprimer la propriété d' « avoir un sens », ou de l'intentionalité.

Il n'est donc pas étonnant que du traitement du problème de l'intentionalité dépende pour une bonne part la crédibilité des sciences cognitives. En vertu de quoi un système computationnel — comme l'est selon l'hypothèse cognitiviste l'esprit humain — représente-t-il le monde ? Nous allons voir que pour répondre à cette question, il est inévitable de prendre position sur d'autres problèmes, et en particulier sur l'identification des contenus mentaux (les conditions auxquelles deux contenus mentaux sont des instances d'un seul et même type), c'est-à-dire sur la taxinomie des contenus, et sur la validité de la thèse de la dépendance systématique (*supervenience*) des états mentaux par rapport à des états cérébraux[2]. Les divergences apparaissent dès que l'on pose les conditions auxquelles il est possible de parler d' « intentionalité », de quoi une telle intentionalité doit être faite, ou, en d'autres termes, sur quel terrain il convient de se situer pour en déterminer les composantes. De la grande variété des réponses, émergent deux traditions. La première rassemble de nombreux penseurs qui appartiennent aussi bien à des courants rationalistes qu'empiristes, tels que Descartes, Berkeley, Hume, Kant ou le Carnap de l'*Aufbau* ; elle pose que l'identification des contenus de pensée est méthodologiquement indépendante de leurs causes extérieures. Les *Méditations* de Descartes constituent un exemple canonique de cette démarche, qui consiste à analyser des contenus de conscience accessibles par pure introspection et, de ce fait, points de résistance à

1. Franz BRENTANO, *Psychologie vom empirischen Standpunkt*, 3 vols, Leipzig, Felix Meiner Verlag, 1924-1928, vol. 1, p. 102.

2. Dire qu'un état mental occurrent A d'un sujet donné dépend systématiquement d'un état cérébral B, c'est dire qu'il existe toujours un certain état cérébral B qui sert de substrat à A. Si un état mental A' est différent de A, le substrat B' de A' est en vertu de cette thèse nécessairement distinct du substrat B de A. Pour un examen de la thèse de la dépendance systématique défendue par D. Davidson, cf. ENGEL, 1986.

l'effet dévastateur du doute hyperbolique. Quel que soit l'état du monde, et, en particulier, qu'il existe ou non, je n'en ai pas moins l'impression d'être dans cette chambre, de fumer cette pipe, etc.[3].

L'autre école est naturaliste, et s'inspire des leçons de Dewey et de Quine. La psychologie doit de son point de vue s'efforcer de suivre les leçons de la physique et de la biologie : les événements mentaux ne sont analysables que comme les réponses de l'organisme à des états de l'environnement. On ne peut donc expliquer l'intentionalité du mental ni identifier le contenu des états mentaux sans mettre en relation un état mental ou computationnel avec certains états de choses.

Toutes les analyses récentes de l'intentionalité s'efforcent de déterminer la part qui doit être reconnue respectivement à la thèse computationnelle — qui, comme on va le voir, est étroitement solidaire de la première perspective — et au point de vue naturaliste quant à la détermination des contenus des états mentaux. Le présent travail a pour objectif d'éclairer le développement de la pensée de Jerry Fodor concernant la théorie de l'intentionalité. Une telle mise au point n'est sans doute pas superflue, dans un contexte où les examens critiques du cognitivisme fodorien ont souvent la fâcheuse tendance de combattre une théorie qu'il n'a jamais soutenue, théorie que l'on pourrait résumer en une formule : « le sens d'un symbole se réduit à sa syntaxe »[4].

Jerry Fodor estime en 1981 que, d'un certain point de vue, l'opposition entre solipsisme méthodologique à la Descartes et naturalisme à la Dewey est *résolue* dans le cognitivisme, c'est-à-dire dans la thèse selon laquelle les processus mentaux sont de nature computationnelle. La condition de « formalité » pose, en effet, que tout processus mental est formel, en ce sens qu'il n'a accès qu'aux propriétés formelles des représentations. Peu importe que l'information ainsi « traitée » soit produite par l'effet de l'environnement sur les « capteurs » d'un robot, sur la périphérie sensorielle d'un homme ou par les stimulations électriques qui affecteraient directement un cerveau (version contemporaine de l'hypothèse du « malin génie ») ; dans tous les cas, seules peuvent être déterminantes pour le système les propriétés *formelles* des représentations[5]. La

3. Il faut cependant relativiser cette affirmation dans le contexte contemporain de l'analyse du « contenu étroit ». Comme nous le verrons plus bas, un solipsisme ontologique et non pas seulement méthodologique serait incapable de fournir une notion de contenu, faute de pouvoir rendre compte de la relation causale entre le monde et les représentations.

4. Voir, par ex. PUTNAM, 1988, p. 6 sq., ou SEARLE, 1980.

5. Cette condition de formalité, particulièrement contraignante en Intelligence Artificielle, suggère le rapprochement entre la problématique de chercheurs comme Newell avec l'idéalisme transcendantal. Cf. mon article, « L'Intelligence Artificielle comme philosophie », *Le Débat, 47,* nov. déc. 1987, p. 88-102.

condition de formalité paraît donc interdire aux processus mentaux tout accès aux propriétés proprement *sémantiques* des représentations ou des contenus mentaux : cette condition s'accommode très bien du fait que les premières n'aient aucun référent, ou que les seconds soient faux, c'est-à-dire ne correspondent à aucun état de choses extérieur. Fodor prend l'exemple du programme SHRDLU de Winograd : le micromonde de blocs dans lequel le système s'oriente n'existe pas ; « c'est un simple ordinateur qui rêve qu'il est un robot » (Fodor, 1981, p. 232)**.

Cependant cette condition de formalité est mise à l'épreuve par le type d'interprétation des états mentaux qu'elle suggère. Quelques rappels sur ce qui distingue les contextes opaques des contextes transparents permettront ici de montrer pourquoi.

LA TRANSPARENCE

Dans un système extensionnel (comme l'est généralement celui d'une théorie scientifique), les propositions enchâssées sont telles que 1) leurs constituants sont interchangeables *salva veritate* avec des constituants équivalents ; et 2), qu'on peut toujours effectuer une généralisation existentielle. Dans un tel système, je peux par exemple dériver les propositions (3) et (4) des propositions (1) et (2) :

(1) L'étoile du soir est l'étoile du matin.
(2) Il est vrai que Vénus est l'étoile du soir.
(3) Il est vrai que Vénus est l'étoile du matin.
(4) Il est vrai qu'il existe quelque chose du nom de Vénus.

Or si l'on veut édifier une psychologie des états mentaux destinée à rendre compte du rapport entre état mental et comportement, comme cherche à le faire le psychologue, qu'il soit fonctionnaliste au sens étroit du terme ou naturaliste, on doit bien reconnaître qu'aucune interprétation transparente ne rendra justice aux attitudes propositionnelles qui sont celles des sujets. C'est l'opacité des contenus d'attitudes propositionnelles qui, aux yeux de Fodor (1981), constitue un argument en faveur de l'hypothèse computationnelle.

** Pour plus de précisions concernant les références placées entre parenthèses dans cet article se reporter à la Bibliographie, p. 32.

L'OPACITÉ

Si l'on s'intéresse par exemple à la croyance d'un sujet X relative aux propositions enchâssées 2-4, on peut très bien décrire dans les termes utilisés par X [6] un ensemble de croyances propres à X tel que :

(5) X croit que Vénus est l'étoile du matin.
(6) X croit que Vénus n'est pas l'étoile du soir.

En outre, du simple fait qu'un sujet croie une certaine proposition, dont l'expression *de dicto* est, par exemple,

(7) X croit que les extra-terrestres sont bien disposés à l'égard des humains,

on ne peut dériver

(8) Il existe des extra-terrestres.

Cette impossibilité caractéristique de pratiquer dans le contexte des attributions de croyances *de dicto* n'importe quelle substitution de termes équivalents (c'est-à-dire de même extension) ou de dériver une proposition existentielle non enchâssée vraie — en d'autres termes, cette opacité des attributions de croyance *de dicto* — est ce qui rend le solipsisme méthodologique si séduisant en psychologie :

> « les attributions opaques sont vraies en vertu de la manière dont l'agent se représente *à lui-même* l'objet de ses désirs (intentions, croyances, etc.). Et, par hypothèse, ce sont ces représentations qui ont un rôle dans la causation des comportements de l'agent » (*ibid.*, p. 235).

SEMI-TRANSPARENCE OU OPACITÉ COMPLÈTE ?

Le problème du contenu des états mentaux est cependant compliqué du fait que si la condition de formalité était pleinement adéquate, on ne

6. Il est également possible de décrire les croyances de X *« de re »*, c'est-à-dire dans les termes de celui qui rapporte les croyances de X.

devrait pas rencontrer de cas où l'opacité est en quelque sorte incomplète, c'est-à-dire où l'attribution de croyance fait intervenir des conditions intrinsèquement référentielles. Or ces cas se rencontrent, comme le rappelle Fodor lui-même, tout particulièrement dans les contextes où les attributions de croyances font intervenir des déictiques. On est parfois conduit, par exemple, à considérer comme identiques des contenus formellement distincts mais coréférentiels ; ainsi si Pierre et moi croyons l'un et l'autre que je suis malade, la formule du langage de la pensée qui exprime ma croyance est

(9) je suis malade,

tandis que celle du langage interne de Pierre est

(10) elle est malade.

D'autre part, en raison du fait que les pensées démonstratives supposent l'existence d'un référent, on ne peut décrire à l'aide d'un démonstratif une pensée démonstrative qui serait dépourvue de référent, telle que

(11) Jean croit que c'est un OVNI,

si « c' » n'a aucun contenu. Le contenu strictement individuel d'une pensée est donc ici insuffisant pour rapporter le contenu de croyance[7].

Réciproquement, la pensée que l'on peut exprimer en une certaine occurrence en disant :

(12) Je trouve qu'on est bien ici,

doit pouvoir être la même pensée, survenant dans un autre lieu, que celle que l'on exprime en disant :

(13) Je trouve qu'on est bien là-bas.

Ces remarques conduisent Fodor à découvrir les limites du principe de formalité en tant qu'il devrait conduire à une taxinomie des états mentaux *entièrement opaque*. Une telle taxinomie ne parviendrait pas à recon-

7. L'idée que l'expression d'une pensée dont une composante n'a pas de référent échoue à représenter une pensée déterminée, et donc à avoir un contenu, est défendue par des auteurs tels que G. Evans et J. McDowell (cf. EVANS, 1982). D'autres auteurs, comme PERRY, 1977 et 1979, défendent en revanche l'idée que des énoncés de ce genre ont un contenu.

naître l'identité de type entre des contenus comme (9) et (10), ou (12) et (13). Il ne faut pourtant pas abandonner l'opacité, qui s'impose pour les raisons évoquées plus haut. On doit simplement reconnaître l'existence d'une tension entre l'approche fonctionnelle de l'attribution des croyances et l'exigence sémantique minimale qu'impose l'identification de certains contenus. Tension qui exige seulement, du point de vue qu'exprime Fodor dans cet article, quelques aménagements ou du moins un *modus vivendi* :

> « si l'on construit une taxinomie de manière *purement* formelle, on a une identité de croyance qui va de pair avec une différence de valeur de vérité. D'un autre côté, si on l'élabore à partir de critères préthéoriques, on n'arrive plus à comprendre que ce sont les croyances et les désirs qui font agir les gens » (*ibid.*, p. 238-239).

Car une fois perdue la condition de formalité, on ne peut plus comprendre l'efficacité causale des processus mentaux, on ne peut plus comprendre pourquoi les états mentaux et les comportements s'enchaînent d'une façon qui est descriptible comme une dérivation.

L'article de 1981 se conclut sur un bilan nuancé. Il y a convergence entre une taxinomie des contenus faisant droit à leur opacité et la condition de formalité — tandis que la taxinomie transparente qu'exigerait une psychologie naturaliste est incompatible avec cette condition. Cependant, il ne faut pas confondre ce résultat avec la liquidation de toute ambition naturaliste ; si l'on a montré la plausibilité de l'idée que les *opérations mentales* ne sont sensibles qu'à la forme des symboles manipulés, ou, en d'autres termes, n'ont pas d'accès à la *sémantique* de ces symboles, *il ne faut pas en conclure que les représentations n'ont pas de propriétés sémantiques.*

Tout en étant convaincu de l'intérêt que représenterait une théorie naturaliste du sens, Fodor est convaincu qu'elle est hors d'atteinte ; elle exigerait, en effet, que soit disponible une science universelle achevée, nous permettant de donner *de façon nomologique* les relations causales entre les référents et les représentations mentales telles qu'elles sont établies par la connaissance des référents qu'atteint chaque science particulière : « on ne peut pas faire de psychologie naturaliste de la référence sans avoir une façon de dire ce qu'*est* le sel ; laquelle de ses propriétés détermine ses relations causales » (*ibid.*, p. 250). C'est pourquoi « une psychologie naturaliste reste une sorte d'idéal de la raison pure » (*ibid.*, p. 252).

Les années 1980 ont vu s'amplifier le débat sur le « site » de l'intentionalité, des auteurs tels que Tyler Burge montrant l'incapacité de la thèse solipsiste méthodologique soutenue par Fodor — laquelle identifie les

contenus à des états internes particuliers de traitement de l'information — à rendre compte de certains types d'attributions de croyance assez proches du cas de l'emploi du mot « eau » sur la Terre Jumelle inventé par Putnam[8].

Ce que montrait Putnam dans son célèbre article, c'est que l'extension des mots de la langue naturelle, tels que le terme d'« eau », n'est pas fonction de l'état psychologique du locuteur. Par conséquent, si l'on admet que connaître le sens d'un mot consiste à être dans un certain état psychologique, on ne peut soutenir que le sens d'un terme détermine son extension, comme on le dit habituellement dans une interprétation libre, psychologisée, du célèbre texte de Frege (1971)[9].

Imaginons, en effet, qu'il existe une autre terre entièrement semblable à la nôtre dans ses moindres détails, y compris en particulier dans le fait que tout terrien a une contrepartie, un « Doppelgänger » entièrement identique à lui quant à ses états cérébraux, sur terre jumelle, à l'exception d'un seul fait : la substance nommée « eau » sur terre jumelle, et qui a toutes les caractéristiques phénoménologiques de ce que l'on appelle « eau » sur terre, a en réalité sur terre jumelle une composition chimique différente, XYZ. Supposons, en outre, que deux habitants jumeaux, nommés respectivement par nous Oscar 1 et Oscar 2, pensent tous deux que « l'eau du bain est trop chaude », et qu'ils aient cette pensée à une époque où l'avancement de la chimie est tel que ni l'un ni l'autre ne disposent encore des moyens de distinguer la composition moléculaire des deux types de liquides. Dirons-nous qu'ils ont la même pensée ?

Lors même que l'argument de Putnam concernait le sens des mots de la langue naturelle, il est clair qu'il s'applique aussi aux expressions du langage de la pensée : les deux pensées exprimées par deux formules du « mentalais » de l'un et de l'autre ne peuvent pas être identiques puisque les conditions de satisfaction des concepts exprimés par le mot d' « eau » ne sont pas les mêmes sur terre et sur terre jumelle, lors même que les deux sujets sont dans un état cérébral identique.

L'erreur pour Putnam remonte à une certaine interprétation psychologiste de Frege, erreur qui constitue une tentation permanente pour le fonctionnaliste : elle réside dans le fait de penser que connaître le sens d'un terme consiste à être dans un certain état psychologique, *et* que le sens d'un terme détermine son extension. Mais la leçon de l'expérience de pensée de Putnam va plus loin que la thèse selon laquelle le sens des mots de la langue naturelle « ne sont pas dans la tête » : elle remet

8. Cf. PUTNAM, 1975.

9. Sur la psychologisation des réflexions frégéennes sur le sens et sa portée dans la philosophie contemporaine du langage, cf. PROUST, 1981.

principalement en cause le bien-fondé du solipsisme méthodologique, c'est-à-dire la doctrine selon laquelle, d'après les termes de Putnam, « aucun état psychologique proprement dit ne présuppose l'existence d'autre individu que celui auquel l'état psychologique est attribué » (Putnam, 1975, p. 220), en particulier si cette doctrine a l'ambition de livrer une théorie du contenu. On peut donc tirer de cette expérience de pensée la conclusion qu'aucune théorie des états mentaux à base individualiste ne pourra livrer une théorie de l'intentionalité de ces états.

Burge tire une morale sensiblement différente de l'article de Putnam [10]. Ce qui permet de spécifier le contenu mental de quelqu'un, c'est le sens des occurrences enchâssées, « obliques », dans des phrases telles que « Jean croit que l'eau du bain est chaude ». Si le contenu mental d'Oscar 1 et d'Oscar 2 diffèrent, c'est simplement parce que la pensée de l'un concerne de l'eau, tandis que celle de l'autre ne renvoie pas en fait à de l'eau. Il est donc impossible, du point de vue de Burge, d'invoquer comme le fait Putnam la déicticité cachée de termes de substance, dont l'extension changerait en fonction du contexte. Burge montre que l'on ne peut attribuer au concept EAU exprimé par le mot d'« eau » le changement d'extension du mot « ici » (dans la reformulation proposée par Putnam : dire que *x* est de l'eau, c'est dire que *x* est identique au liquide qu'on appelle « eau » ici) parce que dans ce cas, XYZ serait de l'eau, ce qui est faux.

Renonçons donc à chercher une solution en termes de l'indexicalité cachée des termes de substance : la différence entre les deux emplois du mot « eau » sur Terre et sur Terre Jumelle n'est pas telle qu'elle affecterait les extensions d'un sens linguistique constant ; *car les deux mots d'« eau » ne partagent même pas leur sens linguistique acontextuel.*

Il est, en outre, du point de vue de Burge sinon faux, du moins équivoque de dire que le sens « étroit » d'un terme comme « eau » ne fixe pas l'extension du mot « eau » comme on est tenté de le dire par suite de l'argument de Putnam (par « sens étroit », on entend le rôle fonctionnel d'une représentation ou d'une phrase du langage de la pensée, le « sens large » désignant ses conditions de vérité). On ne peut évidemment conclure directement d'un contenu de croyance pris *de dicto* — sans parler ici de l'attribution à autrui d'une croyance — aux extensions des termes impliqués dans le rapport de croyance correspondant. Mais d'un point de vue purement sémantique, c'est une vérité nécessaire que « eau » fasse référence à l'eau et seulement à elle.

Puisqu'on ne peut pas spécifier le contenu mental d'un sujet sur une

10. Cf. BURGE, 1982, p. 102 sq.

base individualiste, la stratégie de recherche de Fodor dite du « solipsisme méthodologique » se voit enfermée dans des limites assez étroites : elle ne peut pas prétendre comme le pensait Fodor dans son article de 1981 spécifier les attributions de croyance non transparentes (même les attributions de croyance *de dicto* supposent l'existence d'autres entités), ni de façon générale fournir une théorie de l'intentionalité proprement dite.

Fodor (1987) présente une réponse élaborée aux arguments externalistes. D'un côté, il tire d'une réflexion générale sur le raisonnement causal dans la formation des concepts scientifiques l'idée que l'individualisme méthodologique — à distinguer, comme on va le voir bientôt, du solipsisme méthodologique — est rationnellement justifié dans tous les cas, même quand il s'agit de domaines typiquement relationnels comme l'est en l'occurrence celui de la référence.

De l'autre, il propose son propre diagnostic du problème des terres jumelles, diagnostic qui situe la vraie difficulté non pas dans le rapport entre des intensions « dans la tête » et des extensions qui, précisément, resteraient mentalement indifférenciées, mais dans l'impossibilité où nous sommes par principe de spécifier les contenus étroits (du mentalais) dans la langue naturelle (c'est-à-dire sans « toujours déjà » les ancrer dans un contexte). Revenons sur ces deux types d'arguments.

L'argument méthodologique de fond, auquel Fodor a recours[11], invoque contre l'argument de Putnam-Burge le concept de pertinence causale qui, selon lui, caractérise nécessairement toute taxinomie scientifique. En bref, Fodor distingue de façon très nette deux hypothèses qui étaient restées jusqu'alors inextricablement confondues ; le solipsisme méthodologique et l'individualisme méthodologique. Le premier est une théorie *empirique* sur l'esprit, qui en pose la nature computationnelle. Le second est une règle générale de méthodologie scientifique : les catégories scientifiquement pertinentes sont celles qui permettent des généralisations causales.

Le fait qu'Oscar 1 fasse référence à H_2O tandis qu'Oscar 2 fait référence à XYZ n'est fonctionnellement d'aucune conséquence quant à la suite des attitudes propositionnelles ayant un rapport avec le liquide que l'un et l'autre appellent de l'eau. Négliger cette différence ne constitue pas de ce fait une lacune de la théorie ; c'est une mise en application d'un principe universellement reconnu dans les sciences : n'est théoriquement pertinente qu'une propriété ayant des implications causales. Par exemple,

11. Cf. FODOR, 1987, p. 42-43.

deux prédicats parfaitement bien définis en termes purement physiques tels que

« être une particule P au temps t », et
« être une particule F au temps t »,

qui s'appliquent l'un et l'autre à toute particule physique telle respectivement que, au temps t, ma pièce de 1 F est tombée sur pile (P) ou sur face (F), ne peuvent évidemment permettre aucune généralisation causale, et sont donc pour cela dénués de tout intérêt théorique. Certaines catégories scientifiques pertinentes peuvent en revanche, sans paradoxe, être relationnelles et individualistes : elles sont individualistes dans la mesure où elles sélectionnent seulement celles des propriétés relationnelles qui jouent un rôle causal. La propriété « être une planète », par exemple, est une propriété relationnelle mais qui est déterminée de manière purement causale.

Résumons-nous : si c'est l'explication causale du comportement qui fait l'objet de l'investigation psychologique, il n'y a aucune raison de renoncer à dire que l'état mental qui s'exprime chez les deux Oscars par la formule du langage de la pensée correspondant à « l'eau du bain est trop chaude » est identique, même si l'on s'écarte en cela des intuitions du sens commun. L'identité du contenu étroit de cet état mental correspond au fait que les mêmes types d'inférences seront tirés par l'un et l'autre de la pensée en question, et conditionneront des comportements identiques [12].

La première partie de la réponse de Fodor consiste ainsi à dire que toute « individuation » *dans les sciences* est de type « individualiste », l'individualisme méthodologique soutenant dans le cas de la psychologie que les états mentaux sont individués par leurs capacités causales ; et à renvoyer à Burge l'ascenseur causal : à quoi peut bien servir une théorie psychologique externaliste qui découvre des différences de contenu mental *sans contrepartie* dans les états cérébraux : « comment les différences de contexte pourraient-elles affecter les capacités causales des états mentaux d'un sujet sans affecter l'état de son cerveau ? » (Fodor, 1987, p. 41-42). On ne peut évidemment abandonner la dépendance systématique du mental sur le cérébral — ce qui serait l'une des issues possibles hors du labyrinthe de Putnam-Burge et qu'emprunte précisément Burge — sans en même temps se priver du moyen de rendre compte de la causation mentale.

Venons-en alors au diagnostic que Fodor propose, diagnostic destiné

12. *Ibid.*, p. 34.

à préserver à la fois la dépendance systématique du mental à l'égard du cérébral *et* le lien entre contenus et extensions. « Les exemples de Terre Jumelle ne suppriment pas le lien entre contenu et extension ; ils le relativisent seulement au contexte » (Fodor, 1987, p. 47). En d'autres termes, il suffit de dire que mon jumeau et moi partageons le même contenu étroit de croyance, désir, etc. Mais, dans le contexte de la Terre, l'extension déterminée par ce contenu est H_2O, tandis que dans le contexte de la Terre-Jumelle, c'est XYZ, que la pensée concernée soit celle de mon jumeau ou la mienne. Deux pensées sont donc de même contenu à la seule condition que leurs conditions de vérité coïncident pour un contexte donné.

On reconnaît dans ce diagnostic la solution *de dicto* écartée par Putnam dans l'article princeps selon laquelle « l'eau est ce qui est semblable à ce qu'on appelle "eau" ici », puis rejetée, pour des raisons différentes, par Burge. La difficulté majeure de cette solution est que le contenu étroit paraît à peine mériter le nom de contenu dans la mesure où il n'est *pas encore sémantiquement évaluable,* puisqu'un contexte n'est pas encore donné qui rende possible cette évaluation. Par opposition à un tel contenu étroit, ce qu'on appelle contenu dans la réflexion sémantique traditionnelle comme la pensée frégéenne, la proposition en soi de Bolzano ou la phrase éternelle de Quine ont une valeur de vérité déterminée en ce sens qu'elles *incluent* les déterminants contextuels du sens. Il est difficile de voir dans le contenu étroit autre chose qu'un déterminant du contenu, clairement inspiré du « caractère » de Kaplan [13], déterminant qui ne fournit pas une condition suffisante mais seulement dans le meilleur des cas une condition nécessaire de l'intentionalité.

Cette difficulté se double du fait que le contenu étroit, comme on l'a vu, n'est jamais spécifiable dans la langue. C'est là aux yeux de Fodor un fait empirique : « le contenu que la phrase anglaise exprime est *ipso facto* un contenu *ancré,* par conséquent *ipso facto* un contenu qui *n'est pas* étroit » (Fodor, 1987, p. 50). La difficulté se précise alors de la manière suivante : qu'est-ce qui autorise ici le théoricien à opposer un contenu seulement « potentiel » à un contenu « en acte » ? Comment la relation de référence (d'un mot à ce qu'il dénote) peut-elle s'articuler sur une relation potentielle de sens (d'un mot à ce qu'il signifie) dans une théorie *naturaliste* des contenus (entreprise parfaitement étrangère à Frege et à

13. Cf. KAPLAN, 1989. Rappelons que Kaplan distingue le « caractère », comme fonction fixée par des conventions linguistiques d'un contexte à un contenu, du contenu qui est une fonction des circonstances d'évaluation à une extension appropriée (cf. *ibid.*, p. 500-507). Par exemple, le mot « je » a pour caractère ' « je » fait référence à l'agent dans le contexte considéré ', et a pour contenu dans les présentes circonstances l'auteur de cet article, J. P.

Bolzano et qui a conduit Quine, comme on le sait, à renoncer aux contenus) ? Comment, enfin, est-on assuré que le « contenu étroit » constitue *déjà* un contenu, par opposition à une simple forme ? Ce contenu « en forme de caractère » *(characterlike)* est-il autre chose qu'une combinaison (que la tradition aurait jugée « nominale ») de jetons symboliques pourvus d'une forme et d'une fonction ? N'est-ce pas une pétition de principe que d'y lire les prémices d'un contenu [14] ?

On peut être tenté de donner raison à Fodor lorsqu'il reconnaît que « la théorie qui est ici en train d'émerger est, en un sens, une théorie "sans contenu" du contenu étroit », mais hésiter à reconnaître avec lui qu'il s'agisse encore « d'une théorie pleinement intentionnaliste » (Fodor, 1987, p. 53). Il faut cependant se souvenir que la théorie du contenu comporte désormais deux niveaux, et chercher dans la suite de l'ouvrage la seconde partie de la théorie, celle qui doit permettre d'apporter une réponse aux questions laissées en suspens, telle que celle de l'évaluation sémantique des attitudes propositionnelles ou celle du lien entre contenu « étroit » et contenu « large ». Le premier niveau, dit du « contenu étroit », est solidaire de la version fodorienne de la théorie représentationnelle de l'esprit, en vertu de laquelle les états et les processus mentaux sont computationnels, en ce sens que les représentations obéissent à des règles formelles de combinaison, et reçoivent leurs propriétés causales en partie en vertu de leurs propriétés formelles. Le contenu étroit est ainsi requis, comme on l'a vu, pour garantir le caractère causal des relations entre états mentaux. Reste alors à comprendre en naturaliste ce qui est en jeu dans l' « interprétation d'un symbole primitif » — non logique — « du mentalais dans un contexte donné » (Fodor, 1987, p. 98).

La théorie causale de la référence paraît ici fournir le cadre général de la solution recherchée. De façon générale, une telle théorie explique la relation de référence entre une expression et une entité par l'existence d'une relation converse de causalité entre l'objet ou la propriété dénotés et le nom propre ou le prédicat qui les dénotent ; initialement destinée à rendre compte de la référence des expressions du langage naturel, cette théorie peut aussi bien s'appliquer aux symboles du langage de la pensée.

En fait, une théorie causale est notoirement trop sommaire pour rendre compte du contenu des états mentaux. Comme le montre Dretske (1981), il convient de distinguer entre causalité et régularité nomique pour pouvoir rendre compte du lien informationnel qui existe entre, par exemple, deux séries d'événements. L'existence d'une relation causale entre A et B (une mouche dans le champ de vision d'une grenouille

14. Sur l'opposition traditionnelle entre définition nominale et définition réelle, cf. PROUST, 1986, p. 66 sq.

provoque une excitation neuronale particulière qui déclenche à son tour la réponse « happer au vol ») ne suffit pas à dire qu'un flux d'information se produise de A vers B. La théorie informationnelle du sens que suggère le travail de Dretske pose que l'intentionalité présente dans la transmission et la réception de l'information dépend non d'une relation de causalité, mais de « régularités nomiques », c'est-à-dire de corrélations réglées entre événements qui ne sont pas nécessairement déterministes, mais qui sont contrefactuellement stables.

Dans la mesure où elles éclairent la complémentarité du contenu étroit et du contenu large, les analyses de Dretske apportent de l'eau au moulin de Fodor : le *« narrow content »* renvoie à la structure cognitive d'un concept (c'est-à-dire à ses effets et conséquences fonctionnels), tandis que le *« wide content »* renvoie aux origines informationnelles des concepts. Dretske admet ainsi à la suite de Putnam que l'on peut parfaitement posséder un concept sans connaître les conditions nécessaires et suffisantes de son application. L'apprentissage du sens d'un mot se fait par exposition à des signaux porteurs d'information, laquelle dépend des régularités physiques de l'environnement, et non pas des caractéristiques physiques isolées des expériences d'un sujet. Ce sont ces régularités qui décident du fait que le concept d' « eau », par exemple, ait le contenu informationnel qu'il a. Pour le comprendre, revenons à l'exemple de Putnam. Imaginons maintenant que l'on compare le contenu informationnel de deux « presque-jumeaux », le terrien Oscar 1 et un non-terrien jumeau Oscar 3, qui ont été exposés à deux types de régularités : le premier est entouré d'eau dont la composition moléculaire est H_2O, tandis que le second a appris que l'eau avait deux types de composition moléculaire possibles : H_2O ou XYZ. Imaginons enfin que, par hasard, et à son insu, toutes les expériences de l'eau qui ont été faites par Oscar 3 (le non-terrien) étaient des expériences d'H_2O. Le fait que rien ne permette de distinguer les échantillons d'eau — qu'il s'agisse en fait dans les divers cas rencontrés par nos deux sujets de la substance nommée « eau » sur terre — n'empêche pas qu'ils aient une valeur informationnelle différente ; cette différence vient de la différence entre les types d'information auxquels les sujets ont été exposés pendant leur période d'apprentissage, différence qui détermine le contenu des concepts en question [15].

15. L'une des difficultés de la thèse défendue par DRETSKE, 1981, réside dans le caractère flou et non rigoureux de la notion de « période d'apprentissage ». Comme l'écrit FODOR, 1987, p. 103 : « Il n'y a pas de moment à partir duquel l'usage que l'on fait d'un symbole cesse d'être simplement en voie de formation et commence à se faire, en quelque sorte, pour de bon. »

LE PROBLÈME DE LA ROBUSTESSE DU SENS

Néanmoins, la solution de Dretske, de même que les autres théories causales ou informationnelles du contenu, n'échappent pas à une série de difficultés évoquées dans les publications les plus récentes de Fodor, difficultés qui sont très largement imputables à ce que Fodor appelle la « robustesse » du sens, c'est-à-dire à la propriété en vertu de laquelle les occurrences d'un symbole peuvent être causées par les moyens les plus divers sans pourtant cesser de signifier une seule et même chose.

La première difficulté que Fodor a remarquée, dans *Psychosemantics,* tient à l'incapacité de ces théories de rendre compte de la *méprise représentationnelle (misrepresentation).* Prenons le cas d'une représentation de « vache » qui est causée par une occurrence de vache, mais qui peut également être causée, dans certaines circonstances (obscurité, brouillard...) par autre chose qu'une vache — par exemple un orignal. Faut-il en conclure que tantôt les occurrences de « vache » sont causées par des vaches, tantôt elles sont causées par des orignaux ? Dans ce cas, le contenu du symbole « vache » est la propriété disjonctive « être soit une vache soit un orignal ». Une telle théorie causale ne peut donc rendre compte de l'erreur de représentation. Elle ne parvient pas à distinguer le cas où le concept représenté est disjonctif (comme « être une vache ou être un orignal ») du cas où le concept de vache a par erreur été appliqué à quelque chose qui ne tombait pas sous lui (cf. Fodor, 1987, p. 102). Le défi qui se présente à une théorie *naturaliste* est de résoudre ce problème de l'erreur — qu'on nomme « problème de la disjonction » — sans s'appuyer circulairement sur des concepts intentionnels, en invoquant par exemple des « circonstances normales », ou « idéales », ou « sélectionnées par l'évolution », etc.

Fodor (1987) propose de résoudre le problème de la disjonction en faisant appel à l'asymétrie — laquelle n'est pas selon toute apparence intentionnelle — de la dépendance entre, respectivement, l'occurrence d'une propriété causant « normalement » le symbole mental de « vache » et ce symbole mental, et l'occurrence de la propriété « orignal » causant « accidentellement » ce même symbole et le symbole. C'est parce que les vaches donnent lieu à la représentation de « vache » que les orignaux conduisent parfois à dire, ou à se représenter « vache », tandis qu'il n'est pas vrai symétriquement que ce soit parce que les orignaux produisent le symbole de « vache » que les vaches conduisent à se représenter « une

vache ». Cette théorie permet d'expliquer l'erreur en termes non intentionnels, empiriques, de dépendance asymétrique entre des relations causales. Selon la formule de Fodor (1988b), « les occurrences fausses dépendent métaphysiquement des vraies » (p. 29).

Cependant, le problème de la disjonction, qui dans *Psychosemantics* est associé à la méprise représentationnelle, est d'application beaucoup plus large, comme Fodor le montre dans des textes encore inédits (1988a, 1988b). Il concerne non seulement l'application erronée des symboles, mais plus généralement *tous les cas où les occurrences symboliques ne sont pas causées par les objets ou propriétés entrant dans l'extension du symbole.* Le problème de la disjonction naît, en fait, précisément de la confusion ou de la distinction insuffisante entre « signifier » et « être porteur d'information ». L'information, dont un symbole est le porteur, est solidaire du lien causal qui s'établit entre une entrée perceptive et une représentation symbolique ; mais le sens est très largement indépendant de ce lien causal. En d'autres termes, à cause différente d'occurrences d'un certain type symbolique, information différente ; en revanche, « le sens d'un symbole fait partie des choses que toutes ses occurrences ont en commun, quelle qu'ait été leur histoire causale » (Fodor, 1988b, p. 28). Le sens a une « robustesse » qui fait défaut à l'information dont un symbole ou un objet, ou un événement peuvent être porteurs.

La généralité de la difficulté apparaît si l'on remarque qu'un problème de disjonction se manifeste dans un type de cas qui n'a rien à voir avec l'erreur : celui où un symbole est produit non par la perception d'un objet (soit dans l'usage où il fonctionne comme étiquette — cette forme appauvrie de langage que Wittgenstein évoque en tête de ses *Recherches*), comme dans le cas précédent (« c'est une vache », ou seulement « vache » !), mais par un autre symbole avec lequel il entretient une certaine relation (comme « une vache est un mammifère » ou bien « il y a des vaches en Suisse »), ou bien, les productions mentales « du coq à l'âne » si l'on peut dire, comme lorsqu'une occurrence mentale de taureau évoque une occurrence de vache.

Fodor étend à cette classe de cas la solution qu'il avait appliquée au cas de la méprise représentationnelle. Si un locuteur utilise, par exemple, dans le langage « augustinien » imaginé par Wittgenstein, l'expression « brique ! » non pas pour rapporter qu'il voit une brique, mais pour en demander une, cet usage est asymétriquement dépendant à l'égard d'un usage en quelque sorte fondamental, usage dans lequel existe une certaine relation causale entre une brique et mon expression « brique ». Ce que l'on exprime en termes de langage naturel vaut tout aussi bien du mentalais : on peut faire l'hypothèse qu'il puisse exister, à ce niveau, des

mécanismes qui sous-tendent les relations asymétriques entre divers usages des représentations.

Cette explication, pour ingénieuse qu'elle soit, pose un certain nombre de difficultés que je dois me contenter d'évoquer brièvement. Mais je commencerai par défendre Fodor contre une objection qu'on ne manquera pas de lui faire, et qui est fréquemment élevée contre toutes les tentatives de naturalisation. La théorie de Fodor ne contient-elle pas une circularité vicieuse, en se donnant le droit d'identifier les contenus — des référents de symboles mentaux tels qu'une vache ou un orignal — et en examinant les diverses situations contrefactuelles où ces référents pourraient causer conjointement ou non les symboles mentaux correspondants ? La réponse à cette question n'est pas aisée. Car en un sens, il y a circularité, puisque la théorie ne cherche pas à réduire les faits, les transformer dans un format où ils cesseraient de constituer un certain découpage des phénomènes que la langue naturelle ou scientifique nous transmet. Mais ce n'est pas parce que la théorisation serait restée en quelque sorte insuffisamment poussée qu'elle reste « au niveau des faits ». C'est qu'elle estime que c'est à ce niveau que se forment les conditions ultimes de l'intelligibilité, c'est-à-dire de l'objectivation scientifiquement effectuable.

Ce n'est pas dire pourtant que la solution de Fodor soit, selon ses termes, parfaitement « kasher » du point de vue naturaliste. L'une des exigences du naturalisme est de fournir une explication qui soit de type extensionnel, c'est-à-dire qui puisse être donnée dans un vocabulaire physicaliste. Or la théorie qui identifie les conditions suffisantes du contenu intentionnel dans l'information + l'asymétrie doit postuler le réalisme des propriétés. Ce réalisme permet entre autres de garantir que l'on puisse avoir des relations réglées entre des propriétés qui n'ont pas de porteur, et de leur appliquer de ce fait les manipulations contrefactuelles requises par la condition d'asymétrie. Par exemple, on a le droit de dire que toutes les occurrences de non-licornes ne peuvent causer une occurrence du concept Licorne que si des licornes causaient l'occurrence du concept Licorne au cas où elles existeraient (cf. Fodor, 1988b, p. 37). Mais on peut douter que la solution proposée soit naturaliste si quelques-uns des problèmes essentiels d'une théorie du contenu sont résolus par le truchement d'une ontologie intensionnelle. Par exemple, la question de l'indétermination de la référence, la question de la traduction, la question du vocabulaire logique et mathématique, la question de la nature des termes primitifs sont prérésolues — à bon compte d'un point de vue naturaliste strict — par une telle théorie qui peut identifier les contenus mentaux à des contenus objectifs transindividuels.

Une seconde limitation du naturalisme de cette solution réside dans la

postulation que l'on doit faire de l'existence de mécanismes régissant les divers emplois d'un symbole mentalais. Ces mécanismes sont une hypothèse qui reste pour le moment purement spéculative ; l'idée qu'il existe un contenu en quelque sorte fondamental des concepts en faveur duquel pourrait jouer la dépendance causale paraît difficilement recevable pour un naturaliste. Entre le fait de la robustesse sémantique et l'explication en termes de mécanismes mentaux sous-jacents qui en est finalement proposée, la théorie paraît se borner à proposer comme une solution ce qui ressemble encore à une reformulation du problème, ou, tout au moins, à un programme de recherche.

Fodor évoque lui-même un second ensemble de difficultés, qui tiennent au fait que la condition qui est présentée pour rendre compte du rapport intentionnel est satisfaite probablement par de nombreuses chaînes causales qui constituent elles aussi des cas de dépendance asymétrique (Fodor, 1988b, p. 55). Ne risque-t-on pas dès lors de tomber dans un « pansémantisme » ? Fodor propose de restreindre la classe des cas indésirables en considérant que la robustesse n'est pas caractéristique des rapports causaux non symboliques. Si l'on suppose que

> « les A comme A causent les B comme B, et que les B comme B causent les C comme C, la loi A → C dépend asymétriquement de la loi B → C [...] La dépendance des C à l'égard des A n'est robuste que s'il y a des C non causés par des A. Mais dans la chaîne causale A → B → C, tous les C causés par B sont aussi causés par A »,

ce qui montre que la relation A → C n'est pas robuste.

Un exemple permet pourtant de se demander si la robustesse n'est pas un phénomène courant dans les relations causales. Si un vent de force 8 fait chavirer un voilier, et si le naufrage entraîne la noyade des occupants, le rapport causal entre la force du vent et la noyade est symétriquement dépendant du rapport causal entre naufrage et noyade : si les organismes humains pouvaient s'oxygéner sous l'eau (c'est-à-dire si le lien causal entre B et C était rompu), la force du vent ne pourrait pas entraîner de noyades. La robustesse fait-elle défaut à cette chaîne causale, comme Fodor le dit à propos du cas général ? Ce serait le cas si la noyade des occupants ne pouvait être imputée à d'autres événements qu'à un vent de force 8. Or on peut imaginer qu'un vent de force 9 ou 10, qu'un raz de marée, qu'un tremblement de terre sous-marin, auraient pu avoir le même résultat. Quant au lien causal B → C, on peut imaginer que l'effet aurait pu être causé par une baignade imprudente, un suicide collectif, etc. On peut donc être justifié par la théorie à dire que la noyade signifie la force 8.

Ce que cet exemple montre accessoirement, c'est que la relation

causale elle-même paraît difficilement caractérisable de façon strictement non intentionnelle. Il paraît clair qu'une situation est causale sous une certaine description, soit en vertu de certains traits sémantiquement pertinents. On dira sans doute que l'épistémologie de la causalité ne doit pas être confondue avec l'ontologie physicaliste : la seconde est donnée dans le réel et non dans le discours. Mais ce que l'exemple précédent permet de soupçonner, c'est que l'intuition des contenus, omniprésente, risque de contaminer les critères naturalistes de l'intentionalité. En termes humiens, il paraît difficile en ce point d'abstraire la causalité de l'habitude, et de l'utiliser comme s'il s'agissait d'une relation qui n'était pas déjà construite sur un certain type de pertinence sémantique. Comme on le voit, l'explication naturaliste du contenu risque bien de rester encore longtemps un « idéal de la raison pure ».

Joëlle PROUST,
C.N.R.S.,
C.R.E.A., École polytechnique, Paris.

BIBLIOGRAPHIE

BURGE (Tyler), 1979, « Individualism and the Mental », eds. P. A. FRENCH, T. E. UEHLING, H. K. WETTSTEIN, Minneapolis, University of Minnesota Press, *Midwest Studies in Philosophy*, vol. 4.

BURGE (Tyler), 1982, « Other Bodies », *in* A. WOODFIELD, ed., *Thought and Object*, Oxford, Clarendon Press.

DRETSKE (Fred I.), 1981, *Knowledge and the Flow of Information*, Cambridge, MIT Press.

ENGEL (Pascal), 1986, « L'anomalie du mental », *Critique, 474*.

EVANS (Gareth), 1982, *Varieties of Reference*, ed. J. MCDOWELL, Oxford, Oxford University Press.

FODOR (Jerry), 1981, *Representations. Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, MIT Press.

FODOR (Jerry), 1987, *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MIT Press.

FODOR (Jerry), 1988a, « Information and Representation », manuscrit.

FODOR (Jerry), 1988b, « A Theory of Content », manuscrit.

FREGE (Gottlob), 1971, « Sens et dénotation », in *Ecrits logiques et philosophiques*, trad. Cl. IMBERT, Paris, Le Seuil.

JACOB (Pierre), 1989, « Le problème du rapport du corps et de l'esprit aujourd'hui ; essai sur les forces et les faiblesses du fonctionnalisme », manuscrit.

JACOB (Pierre), 1990, « Externalism Revisited : Is There such a Thing as Narrow Content ? », *Philosophical Review*, à paraître.

KAPLAN (David), 1989, « Demonstratives », *in* J. ALMOG, J. PERRY, H. WETTSTEIN, eds, *Themes from Kaplan*, Oxford, Oxford University Press.

PERRY (John), 1977, « Frege on Demonstratives », *Philosophical Review*, LXXXVI, p. 474-497.

PERRY (John), 1979, « The Problem of the Essential Indexical », *Noûs*, XIII, p. 3-21.

PROUST (Joëlle), 1981, « Sens frégéen et compréhension de la langue », in *Meaning and Understanding*, Berlin/New York, De Gruyter, p. 304-323.

PROUST (Joëlle), 1986, *Questions de forme. Logique et proposition analytique de Kant à Carnap*, Paris, Fayard.

PUTNAM (Hilary), 1975, « The Meaning of Meaning », in *Mind, Language and Reality, Philosophical Papers*, vol. 2, Cambridge, Cambridge University Press, p. 215-271.

PUTNAM (Hilary), 1988, *Representation and Reality*, Cambridge, MIT Press.

SEARLE (John R.), 1980, « Minds, Brains and Programs », *The Behavioral and Brain Sciences, 3*, p. 417-457.

LES REPRÉSENTATIONS INTERMÉDIAIRES *

L'analyse des comportements humains observables se ramène très souvent : (i) à constituer des *représentations symboliques* organisées en systèmes ; (ii) à formuler des *opérations* sur ces représentations qui ainsi deviennent structurées ; certaines opérations sont alors constitutives d'*algorithmes* et de *stratégies* de traitement et de transformation ; (iii) à *identifier les organes* qui seraient les supports physiques des représentations et des opérations ; (iv) à *déterminer des architectures* qui rendraient effectuable l'exécution des opérations mises en jeu par la modélisation des comportements observés. Les sciences cognitives (auxquelles contribuent des disciplines bien constituées et aussi diverses que les mathématiques, la logique, l'informatique fondamentale et l'Intelligence Artificielle (I.A.), l'épistémologie, la linguistique, la psychologie cognitive, la psychophysiologie, la neuropsychologie et la neurobiologie ainsi que certains secteurs de la pathologie...) se déploient par conséquent entre trois pôles : les comportements observables, les représentations symboliques formelles et les neurosciences cognitives.

Les neurosciences cognitives modélisent les comportements observables en partant d'explorations neurobiologiques précises. L'examen des explorations neurobiologiques devrait apporter des éléments importants qui viendraient enrichir les représentations et les techniques de traitement de l'I.A. Par ailleurs, les systèmes informatiques de « compréhension » des signaux physiques (acoustiques, images) ont pour but de transformer ces signaux en une description symbolique. Les méthodes utilisées relèvent essentiellement de la reconnaissance des formes. On complète cependant les algorithmes par des raisonnements qui font appel aux

---

* L'article reprend et développe une première version parue en langue anglaise sous le titre « Intermediate Representations in the Cognitive Sciences », dans *Semiotica*, 77, 1/3, 1989, p. 121-135.

connaissances relatives au domaine traité (processus de la communication par la parole, processus de vision) et aux connaissances contextuelles.

La linguistique, la psychologie cognitive et l'Intelligence Artificielle utilisent des représentations symboliques comme des systèmes de réécriture, les automates et les langages formels « context-free » ; le lambda-calcul, la logique combinatoire, la déduction naturelle ; les formalismes logiques (logique du premier ordre ; logiques modales ; logiques intentionnelles ; logiques non monotones ; logiques temporelles) ; les représentations des connaissances en sémantique des langues naturelles, en I.A., en psychologie cognitive (théories « componentielles » ; réseaux sémantiques ; phénomènes de la typicalité)... La notion de représentation est centrale en psychologie cognitive : « Les représentations sont des constructions circonstancielles faites dans un contexte particulier et à des fins spécifiques » (Richard, 1990, p. 10**). Nous prendrons dans ce qui suit la notion de représentation dans ce sens : une construction circonstantielle établie dans un environnement et adaptée à des fins.

1. — INTELLIGENCE ARTIFICIELLE ET SCIENCES COGNITIVES

Les rapports entre sciences cognitives et Intelligence Artificielle ne sont pas toujours très explicites (pour une bonne présentation des problèmes, voir J.-G. Ganascia, 1990). Il existe, en fait, deux conceptions de l'Intelligence Artificielle, l'une « faible » et l'autre « forte ».

Pour l'I.A. « faible », l'ordinateur est donc seulement un instrument très puissant pour l'étude de l'esprit (opposé au cerveau). Selon cette première conception, plus technique, les représentations symboliques sont des outils commodes pour scruter et analyser les comportements observables et les fonctions de l'intelligence humaine. En revanche, pour l'I.A. « forte », l'ordinateur convenablement programmé serait véritablement une modélisation théorique de « l'esprit », voire, certains vont jusqu'à l'affirmer, « il est un Esprit » : l'ordinateur, muni de programmes adéquats, « comprend » et passe réellement par des « états cognitifs » internes. L'équivalence entre une machine de Turing universelle et un ordinateur permet, selon certains, le rapprochement entre deux machines

** Pour plus de précisions concernant les références placées entre parenthèses, dans cet article, se reporter à la Bibliographie, p. 55.

de Turing spécifiques, l'homme et l'ordinateur (position classique de Newell et Simon). Selon cette seconde conception, plus philosophique, les représentations symboliques seraient alors constitutives de l'Esprit. On affirme, par conséquent, l'existence d'un niveau d'opérations mentales qui mettent en œuvre des processus de transformation de symboles formels ; ces processus constituent alors l'essence du mental ; les traitements procéduraux déclenchés par ces opérations mentales sont physiquement réalisées, de diverses façons, par les cerveaux, de la même façon qu'un programme peut tourner sur différents matériels informatiques. En acceptant les hypothèses de base de l'I.A. « forte », l'analogie « l'esprit est au cerveau ce que le programme est au matériel » justifie pleinement l'étude de l'esprit indépendamment de tous les résultats de la neurophysiologie.

Le rapprochement entre l'homme et l'ordinateur est souvent effectué au travers de la notion de Système de Traitement de l'Information (S.T.I.). Pour Newell et Simon, la structure d'un S.T.I. est celle d'une machine de Turing universelle. Cette machine comprend un processeur, une mémoire, des récepteurs et des effecteurs destinés à communiquer avec le monde extérieur. Selon S.T.I., la pensée humaine est décrite comme un système de manipulation des symboles. Des psychologues, comme J. F. Le Ny et J. F. Richard, donnent pour tâche à la psychologie cognitive de simuler les opérations principales et les états qui se déroulent dans la tête du sujet humain. Ils affirment en particulier que :

> « pour une partie de son activité psychologique, l'individu humain est bel et bien un dispositif de traitement et de stockage de l'information. La partie de l'activité qui est en cause est précisément celle que l'on appelle activité cognitive. (...) En d'autres termes, nous pensons qu'il existe, partout dans l'univers, de l'information, et sur cette terre (pour le moins) diverses sortes de dispositifs matériels qui traitent et stockent l'information. Certes, parmi ces dispositifs existent des différences importantes : par exemple, les uns sont naturels, les autres artificiels, les uns fonctionnent essentiellement à partir d'une neuro-chimie à base de carbone, les autres (jusqu'ici) essentiellement au silicium ; de plus, les uns sont programmés par l'hérédité, l'apprentissage et le milieu social, les autres le sont ici et maintenant par des informaticiens, et demain peut-être par tout un chacun. Mais ces dispositifs relèvent conceptuellement tous de la catégorie des " traiteurs et conservateurs d'informations " » (Bonnet, Hoc, Tiberghien, 1986, p. 281).

La « compréhension » étant simulée par un programme qui doit construire des représentations sémantiques, la question de la nature du « sens » devient incontournable. Comment représenter le « sens » et la « signification » ? Le « sens » d'un signe s'oppose-t-il à une forme matérielle ou est-il, lui-même, appréhendable comme une certaine forme ?

Pour répondre à ces questions, l'I.A. discute de ces problèmes dans un cadre conceptuel beaucoup trop restreint en opposant simplement le « sens » (« meaning ») aux formes matérielles. Or, la signification se construit toujours par interprétation ; elle prend corps par une mise en relation d'un système particulier de représentation du « sens » avec un système particulier de formes mais cette mise en relation doit être assurée par un « représentateur », c'est-à-dire par un utilisateur des représentations significatives. La notion du « sens » n'est donc pas un absolu, elle est toujours relative aux problèmes traités, aux utilisateurs des représentations, aux tâches et aux finalités. Tant en philosophie du langage qu'en sémiotique, de nombreux travaux, classiques pour la plupart, ont montré combien cette dichotomie sens/forme pouvait être parfois trompeuse et souvent trop appauvrissante. En particulier, lorsqu'on aborde la question des représentations sémantiques, il est impossible d'ignorer les réflexions du logicien G. Frege et surtout celles du sémioticien C. S. Peirce. Par exemple, Peirce, en rompant avec la conception purement dyadique du signe qui était apparue avec l'âge cartésien, considère qu'un signe linguistique n'est pas, comme chez F. de Saussure, une unité à double face, opposant un signifiant à un signifié, mais une unité plus complexe qui suppose : (i) un *Representamen* ; (ii) un *Objet* ; (iii) un *Interprétant* :

> « Un *Representamen* est sujet d'une relation triadique à un second, appelé *Objet,* pour un tiers, appelé *Interprétant,* cette relation triadique étant telle que le *Representamen* détermine son interprétant à se tenir dans la même relation triadique au même objet pour quelque interprétant » (Charles S. Peirce, *Collected Papers*, I, 541).

2. — NIVEAUX DE REPRÉSENTATIONS

Les représentations symboliques dans les sciences cognitives posent des questions sur le statut même du cognitif et sur les *niveaux de représentations* :

> — A quel(s) niveau(x) se situent les représentations cognitives ? Quelles sont les architectures qui organisent les niveaux ?
> — Comment relier les représentations symboliques (impliquées par l'analyse des comportements observables) aux organes neurobiologiques qui sont les supports de ces représentations ?
> — Doit-on concevoir des organes « sub-symboliques » qui ne pourraient pas être décrits par des systèmes symboliques ?
> — Plus généralement, par quel processus relier les représentations symboliques et les neurosciences cognitives et rendre ainsi compatibles les descrip-

tions symboliques de la linguistique, de la psychologie cognitive, de l'Intelligence Artificielle avec les modèles neurophysiologiques ?

On peut essayer de répondre à ces questions en des termes bien différents de ceux qui sont habituellement employés, tout particulièrement par les détracteurs de l'I.A. et des sciences cognitives. Par exemple, H. Dreyfus a adressé, en 1969, un certain nombre de critiques au programme de recherche des sciences de la cognition. Selon lui, l'I.A. et les sciences cognitives « orthodoxes » seraient fondées par quatre postulats (biologique, psychologique, épistémologique et ontologique). Il écrit (c'est nous qui soulignons) :

> « [...] Tout ce qui peut être saisi peut également être exprimé sous forme de relations logiques, ou plus exactement *sous forme de fonctions booléennes*, obéissant à cette algèbre qui régit la façon dont sont reliés entre eux les éléments binaires. [...] Toute information fournie à un ordinateur doit l'être *sous la forme d'éléments binaires* [...]. A un certain niveau [...] (on suppose d'ordinaire que c'est au niveau des neurones), le cerveau traite l'information selon des opérations discrètes, grâce à *quelque équivalent biologique de commutateurs de type oui/non*, [...] les éléments binaires correspondant à des parcelles de sensation. [...] L'ordinateur sert de modèle à l'esprit humain [...]. Cette conception de la pensée (est) envisagée comme un traitement de l'information (de façon à) assimiler l'esprit à un ordinateur. [...] *L'esprit peut être envisagé comme un système opérant sur des éléments binaires d'information*, selon des règles formelles » (Dreyfus, 1984, p. 192-193).

Certes, cette critique est déjà largement dépassée, elle a été affinée considérablement par H. Dreyfus dans d'autres écrits, en particulier dans son analyse critique des systèmes experts. Cette conception de l'informatique reste néanmoins présente sous une forme approchée dans de nombreux débats autour de l'I.A. et autour de l'informatique utilisée dans la modélisation de processus cognitifs. Malgré les réelles difficultés rencontrées par l'I.A. (promesses non tenues, « échecs » retentissants...), on déclare que ce qui sous-tend l'optimisme des chercheurs de l'I.A.,

> « c'est la conviction que le comportement de l'intelligence humaine est le résultat d'un traitement de l'information effectué par une forme de calculateur numérique et que, puisque la nature a réussi par ce moyen à produire un comportement intelligent, une programmation adéquate devrait obtenir le même comportement de la part des machines numériques, soit en imitant la nature, soit en concevant des programmes meilleurs encore que les siens » (Dreyfus, 1984, p. 191).

Il apparaît bien maintenant que les « postulats » qui fonderaient, selon leurs détracteurs, les sciences cognitives dites « orthodoxes » reposent,

semble-t-il, sur une évidente sous-estimation importante de la puissance expressive de formalismes logiques et sur une ignorance relative d'un des concepts les plus profonds de l'informatique, la compilation. Abordons très brièvement ces deux points.

2.1. La logique ne repose pas sur les seules structures et fonctions booléennes.

L'innovation fondamentale du logicien G. Frege et, dans une certaine mesure, de C.S. Peirce aussi, fut l'invention du traitement formel de la qualification à l'aide de variables liées, ce qui a nécessité une analyse profonde, déjà entreprise par Aristote, de la structure constitutive des propositions. Contrairement à G. Boole (1848) qui avait analysé « la pensée » en termes de propositions interreliées entre elles, G. Frege (1893) enrichit considérablement l'analyse logique en proposant des techniques d'analyse qui permettraient de descendre dans la structure prédicative des propositions à l'aide de concepts mathématiques profonds (notion de fonction, entre autres). Frege a réellement constitué des *langages formels autonomes* (comme les langages du premier ordre) en « captant » certains des mécanismes formels mis en jeu par les opérations prédicatives des langues naturelles et par certaines opérations de détermination des termes nominaux (universalisation et particularisation). Il en a résulté, par affinements successifs dus à B. Russell, A. Church, S. C. Kleene, J. Rosser, W. V. O. Quine, le Calcul des prédicats et les langages du premier ordre.

Si l'algèbre de Boole décrit adéquatement à la fois la structure engendrée par les connecteurs propositionnels (ET, OU, SI... ALORS, ...) et le fonctionnement des « commutateurs en oui/non », elle n'est pas le paradigme de toutes les structures algébriques que l'on rencontre en logique. En effet, sans descendre dans l'analyse interne des propositions, il existe aussi d'autres structures logiques que les structures booléennes. Par exemple, l'étude de la négation montre que l'on peut concevoir différents systèmes logiques : la logique positive (sans négation), la logique minimale (où logique de la réfutabilité), la logique intuitionniste (ou logique de l'absurdité), qui donne naissance aux algèbres de Heyting, puis enfin la logique « classique » (ou logique de la négation avec tiers exclu) qui est une algèbre de Boole. Certaines logiques modales (où le " possible " et le " nécessaire " sont des modalités qui sont ajoutées aux jugements assertoriques, exprimés par des expressions apophantiques) sont associées à des structures topologiques (le système S_4 dans la hiérarchie de Lewis).

Des méthodes nouvelles sont également apparues en logique, par exemple l'analyse sémantique interprétative « à la Tarski », la théorie des modèles, la déduction naturelle de Gentzen. Le Calcul des propositions et le Calcul des prédicats ne sont plus, actuellement, les seuls langages étudiés par la logique. Par exemple, la Logique Combinatoire (Schönfinkel, Curry) et le Lambda-calcul (Church) sont des systèmes de calcul qui opèrent sur des « fonctions pensées en soi » ; ces formalismes sont puissants, ils développent des langages « sans variables » ; ils sont capables d'exprimer des notions fort subtiles (analyse logique du paradoxe de Russel, recherche des points fixes dans la récursion, construction d'opérateurs complexes, arithmétique formelle, logique illative...). Des logiques temporelles et modales sont apparues avec A. Prior ; des logiques « intensionnelles » se sont développées avec R. Montague. Pour tenir compte des défauts, des exceptions, on explore maintenant des logiques non monotones.

Même en oubliant l'existence des logiques à plusieurs valeurs et des logiques « floues », on ne peut plus dire que les relations logiques sont réductibles aux seules « fonctions booléennes ». La conception des circuits électroniques n'est pas non plus soumise à une simple algèbre « des commutateurs de type oui/non » mais à des opérations et à des architectures assez complexes (V.L.S.I.). De plus, il est toujours possible d'inventer des formalismes calculatoires destinés à capter formellement les mécanismes de certaines situations non mathématiques en les formalisant pour les présenter ensuite, éventuellement, sous forme de systèmes formels. Pour s'en convaincre, il suffit de se reporter à l'histoire de la logique et des mathématiques (algèbre avec Viète ; géométrie analytique avec Descartes ; géométrie infinitésimale et calcul intégral avec Pascal, Leibniz, Newton ; calcul tensoriel avec Einstein, calcul des distributions avec Schwartz...) ; de la physique (cinématique avec Galilée ; mécanique avec Newton, mécanique quantique...) ; de la linguistique (formalisation de certaines opérations syntaxiques et critères de complexité des représentations syntaxiques avec Chomsky...).

2.2. Les langages de programmation de « haut niveau » sont de moins en moins conditionnés par les structures des organes physiques des machines électroniques qui transforment leurs expressions symboliques en d'autres expressions symboliques.

Lorsque les informaticiens réussirent à s'abstraire, en partie du moins, des structures physiques des machines à traiter l'information, ils ont conçu des langages de programmation de plus en plus souples et de plus

en plus aptes à formuler relativement facilement certaines classes de problèmes. Tous les langages de programmation actuels (encore plus ceux du futur) sont capables de représenter des informations structurées extrêmement complexes bien que les supports électroniques des organes de la machine informatique, qui assure les traitements, soient directement adaptés aux codes binaires. Si 0 et 1 sont des symboles qui représentent toujours deux états physiques distincts des systèmes constitutifs des ordinateurs, les structures des langages de programmation de « haut niveau » (FORTRAN, ALGOL, BASIC, PASCAL, APL, C, LISP, PROLOG, ADDA, SMALLTALK...) n'ont cependant plus rien à avoir avec des suites de 0 et de 1, insérées dans des structures booléennes, et réalisées par des réseaux de « commutateurs en oui/non ».

Deux raisons soutiennent l'indépendance des langages « de haut niveau » par rapport aux structures des organes de traitement, elles sont complémentaires. D'un côté, on conçoit, on définit et on organise la syntaxe, la sémantique et même une partie de la pragmatique des langages informatiques de haut niveau sans aucune référence aux supports physiques. D'un autre côté, un programme, appelé compilateur, est enregistré dans la machine, il doit « traduire » automatiquement tous les textes écrits dans ces langages (les programmes) sous forme de « textes compilés » directement compatibles avec les structures internes de la machine ; ces textes compilés sont alors exécutables *par* et *sur* les organes de la machine.

3. — LE CONCEPT DE COMPILATION

Rappelons très brièvement ce qu'est un *processus de compilation*. C'est un ensemble hiérarchisé de programmes qui assure automatiquement les « traductions » entre les *expressions externes*, accessibles aux utilisateurs, et les *représentations internes*, compatibles avec les structures électroniques d'une machine à traiter des informations (un ordinateur). Le compilateur prend pour entrée un programme-source, écrit dans un langage de haut niveau, et produit comme sortie un équivalent présenté sous forme d'une séquence totalement ordonnée d'instructions directement exécutables par la machine. Cependant, le processus est tellement complexe qu'il n'est pas raisonnable de concevoir globalement le programme et de l'implanter en une seule étape. La stratégie adoptée a consisté à découper le programme général en plusieurs phases successives, ce qui conduit à engendrer des représentations symboliques intermédiaires à chacune des

phases du traitement ; plus précisément, chaque phase prend pour entrée une représentation du programme-source, au niveau i, la transforme en une autre représentation du même programme-source, au niveau i + 1, cette dernière sera soit la représentation finale prête à être exécutée, soit l'entrée de la phase suivante.

3.1. Hiérarchie des niveaux.

Il y a une hiérarchie de niveaux depuis les niveaux inférieurs, supports matériels des opérations jusqu'aux niveaux supérieurs où s'expriment les logiciels.

Le langage machine, ou mieux le code machine, au plus bas niveau, est composé d'un nombre fini de types d'opérations et d'instructions. Tout programme se ramène alors à une combinaison (en général complexe) des types d'opérations et d'instructions du code machine. Rappelons la très belle remarque de D. Hofstadter :

> « [...] Chaque nucléotide contient environ vingt-cinq atomes ; imaginez ce que serait l'écriture de l'A.D.N. atome par atome, pour un petit virus (à plus forte raison pour un être humain) — vous aurez une idée de ce qu'est un programme complexe écrit en langage machine et les difficultés que pose la compréhension de ce qui se passe dans un programme en n'ayant accès qu'à sa description en langage machine » (D. Hofstadter, 1986, p. 324).

Le niveau supérieur au code machine est celui du *code d'assemblage.* Le programme est écrit dans un symbolisme qui devient plus accessible aux humains : chaque instruction du code machine est réécrite par une expression symbolique littérale, les opérandes sont accessibles par des noms et non plus par des adresses ; il y a, en employant une expression de D. Hofstadter, « réunitarisation » des instructions du code machine. La correspondance entre le code machine et le code d'assemblage est cependant pratiquement biunivoque. Cette « réunitarisation » a pour but d'épargner des efforts importants de compréhension sans qu'il y ait eu, en fait, de grands bouleversements conceptuels. Un programme, appelé *assembleur*, traduit automatiquement les expressions du code d'assemblage dans le code de la machine.

Le niveau supérieur au code d'assemblage est un *langage informatique de haut niveau* dans lequel on est capable, à l'intérieur du langage, de définir puis de nommer de nouvelles entités à partir d'anciennes déjà connues. Les opérations de définition et de nomination de nouveaux modules sont intégrées au langage lui-même et c'est cette intégration qui fait la puissance des langages de haut niveau. Ainsi, écrire un programme

dans ce type de langages évolués ne se ramène plus à engendrer une simple combinaison séquentielle des instructions de base mais à exprimer des agencements de modules, définis localement et appelés, dans le programme principal, par leurs noms. Dans la programmation de haut niveau, les instructions de base du code machine ne sont donc plus visibles pour le programmeur, alors que dans le code d'assemblage, à des « réunitarisations » symboliques près, la structure de la machine reste pleinement apparente au programmeur et il doit en tenir compte.

Un langage de haut niveau n'est pas utile si l'on ne sait pas passer du niveau supérieur au niveau des codes d'instructions exécutables par la machine. Le programme, appelé *compilateur*, transforme *globalement les programmes* exprimés dans un langage de haut niveau en une séquence d'instructions du langage machine, cette séquence devenant ainsi exécutable. Le compilateur n'assure plus, comme c'était le cas avec l'assembleur, une simple correspondance biunivoque entre les langages source et cible car *la structure des langages de haut niveau ne reflète pas la structure du code de la machine qui exécute les instructions.* Du fait des définitions internes des modules et de leurs appels par des noms, le compilateur doit alors procéder à des réarrangements importants ; il engendre à partir de la séquence des instructions d'entrée une tout autre séquence (en général beaucoup plus longue) d'instructions du langage d'assemblage. La compilation doit également *tenir compte de l'environnement pragmatique du programme* (les états internes actualisés de la machine, le système opérationnel de la machine...) en contrôlant, entre autres, le jeu des référenciations contextuelles des noms des entités définies au cours du programme et le jeu des passages entre les paramètres formels du module appelé par le programme principal et les valeurs actualisées à une étape précise du déroulement du programme. Il est donc indispensable de s'assurer que le compilateur assure une transformation correcte : le *programme compilé* (c'est-à-dire le programme qui est le résultat de la transformation, effectuée par le compilateur et exprimée dans le code de la machine) *exécute-t-il « vraiment » le programme qui a été pensé, voulu et conçu par le programmeur* ? Le compilateur qui, bien que livrant des résultats — le programme « tourne bien » — est-il amené à prendre des décisions implicites qui resteraient cachées à l'utilisateur du programme, et de ce fait, introduiraient des erreurs qui rendent erronés les résultats obtenus ? Le compilateur est-il *fidèle et transparent* ou bien trahit-il l'utilisateur des programmes soumis à la compilation ? Or, si la traduction entre deux langues naturelles n'est pas un processus simple (les textes sources et cibles sont toujours dans une relation d'équivalence, plus ou moins forte, mais jamais dans une relation d'identité), le compilateur doit engendrer, pour sa part, une « équiva-

lence forte » entre un programme source et le programme cible : ces deux programmes doivent *calculer les mêmes fonctions*. Il s'agit donc d'une *équivalence extensionnelle* au sens suivant :

> désignons par P le programme écrit dans un langage LH de haut niveau et par *comp*(P) sa version compilée exprimée dans le code machine CM : si le compilateur est correct, alors P et *comp*(P) sont *extensionnellement* équivalents lorsqu'ils associent exactement les mêmes couples de données et de résultats, c'est-à-dire lorsqu'ils calculent *extensionnellement* les mêmes fonctions.

On sait que, dans la pratique, tous les compilateurs d'un langage ne sont pas nécessairement fidèles et corrects ; certains compilateurs engendrent, au cours du processus de compilation, des vermines ou des petites bêtes conceptuelles (en anglais des « bugs ») qui viennent fausser la « traduction » et produire des résultats suspects.

Un compilateur, en tant que programme enregistré, doit être écrit en général dans un langage de plus bas niveau (en langage d'assemblage ou en langage machine). On a cependant montré que l'on pouvait se servir d'un noyau du compilateur pour étendre le compilateur : on se sert du langage pour étendre le langage.

3.2. Compilation classique des langages de programmation de haut niveau.

Pour un langage de programmation du type de FORTRAN ou ALGOL, nous avons l'enchaînement suivant des phases :

Programme-source → 1. Analyse lexicale → 2. Analyse syntaxique → 3. Génération d'un code intermédiaire → 4. Optimisation du code → 5. Génération du code machine → Programme-cible.

La phase d'analyse lexicale a pour but de reconnaître les catégories morphologiques des symboles du programme-source P. La sortie de cette phase est une représentation $R_1(P)$ où sont indiquées les catégories de chaque occurrence des symboles dans P. La phase suivante est une analyse syntaxique. Elle a pour but de regrouper ensemble des séquences de symboles qui appartiennent à des constituants communs. Le programme d'analyse, lorsqu'il examine la représentation $R_1(P)$, construit des arbres syntaxiques « locaux » qui sont progressivement regroupés dans un arbre « global » qui sera la représentation syntaxique $R_2(P)$ du programme-source P. La phase suivante prend pour entrée l'arbre $R_2(P)$ et lui associe une séquence d'instructions symboliques écrites dans un langage symbolique intermédiaire (appelé « code intermédiaire ») dont

les expressions élémentaires sont des instructions qui mettent en œuvre des opérateurs et leurs opérandes. La séquence d'instructions obtenue à l'issue de cette phase constitue une autre représentation $R_3(P)$ du programme-source ; elle est encore relativement indépendante de la structure de la machine, en tout cas, elle est complètement indépendante des *emplacements physiques* où seront exécutées les instructions (les registres de calcul de la machine) et où seront *stockées les informations* qui y sont contenues (les cellules de la mémoire de la machine). Ce code intermédiaire doit être distingué du code d'assemblage, écrit dans un langage symbolique, qui, lui, devra spécifier exactement les registres dans lesquels devront être exécutées les instructions du programme.

La phase d'optimisation a pour but d'accélérer l'exécution ultérieure du programme en tenant compte des contraintes de l'espace du calcul et en évitant, quand cela est possible, d'exécuter plusieurs fois le même calcul. Cette phase peut être omise sans changer les résultats de l'exécution de P. Pour optimiser, elle transforme la représentation $R_3(P)$ dans une autre représentation $R_4(P)$, écrite dans le même code intermédiaire. Cette nouvelle représentation est destinée à : (i) accroître la vitesse d'exécution (en n'exécutant, par exemple, qu'une seule fois un sous-programme qui est appelé plusieurs fois de suite, avec les mêmes données d'entrée, au cours de l'exécution du programme principal) et à (ii) occuper moins de place en mémoire (en stockant dans un seul emplacement de la mémoire des données et des séquences d'instructions qui apparaissent à plusieurs endroits dans l'arbre syntaxique $R_2(P)$ et dans la représentation $R_3(P)$).

La phase finale, appelée « génération du code machine », engendre, à partir de $R_4(P)$ — ou de $R_3(P)$ —, le code objet du programme qui est une autre représentation $R_5(P)$ du programme-source P. Cette dernière représentation doit spécifier exactement les emplacements des données dans la mémoire, sélectionner les moyens d'accès aux données, sélectionner les registres de calcul où les opérations devront être exécutées, spécifier les emplacements où devront être rangés les résultats (intermédiaires ou finaux) obtenus dans les registres, sélectionner les organes de lecture (organes d'entrée) et les organes d'écriture ou de visualisation (organes de sortie)... Il est clair que la représentation finale $R_5(P)$ est étroitement dépendante de la structure physique de la machine. Elle détermine l'exécution du programme P *par* et *dans* la machine. Chaque changement de type de machine entraînera un changement du code machine et la phase finale du compilateur — en général les phases d'optimisation et de génération de code aussi — devra être réécrite, de façon à s'adapter exactement à la structure de la nouvelle machine.

D'autres représentations partielles et des informations (sur les types de

données, par exemple) sont engendrées et conservées en mémoire durant la compilation. Le compilateur émet aussi des messages au cours de la compilation lorsqu'il détecte des erreurs (lexicales, syntaxiques ou autres) ou lorsque le programme rencontre une difficulté qu'il est incapable de diagnostiquer exactement. Des programmes de correction automatique de certaines erreurs détectées peuvent être adjoints au programme de compilation.

Alors qu'un compilateur transforme globalement un programme, écrit dans un langage évolué, dans un autre programme, écrit dans le code d'une machine, ce dernier programme devenant exécutable dès qu'on lui en donnera l'ordre, un *interpréteur* est un programme qui transforme chaque instruction du programme en une expression du code — on dit que l'instruction est interprétée — pour l'exécuter immédiatement et interpréter l'instruction suivante. Un compilateur est beaucoup plus complexe à écrire mais, en revanche, il permet une meilleure optimisation et une plus grande rapidité d'exécution qu'un interpréteur.

3.3. Compilation, convivialité et linguistique.

L'idée de compilation, ainsi décrite, est née à la confluence des deux grands courants. L'un, plus pratique, est né du désir de rendre plus conviviaux les premiers langages de programmation, trop liés, pour les premiers essais de programmation, aux structures des machines ; l'autre, plus théorique, tire son origine dans la linguistique et dans les études mathématiques qu'elle a suscitées.

En effet, si l'on voulait rendre utilisables la programmation des machines pour exécuter des calculs fastidieux pour des mathématiciens, des astronomes, des physiciens, des ingénieurs, des gestionnaires et ne pas la laisser aux mains de quelques programmateurs spécialisés dans un langage spécifique à un type de machine, il fallait envisager des langages qui, de par la forme de leurs expressions, ne seraient pas trop éloignés des habitudes linguistiques des milieux scientifiques. De plus, les premiers programmes n'étaient pas encore très complexes, ils nécessitaient un savoir technique important (apprentissage du code spécialisé des machines), ils n'étaient pas transparents (un programme écrit dans le code machine ne fait pas apparaître la structure des données manipulées et la logique de l'algorithme) ; ils étaient donc longs à écrire, longs à mettre au point, pratiquement incommunicables à d'autres humains. Comme les calculs scientifiques étaient essentiellement visés par les premiers travaux de programmation, il devenait évident que les langages de programmation devaient ressembler aux langages des mathématiciens

et des ingénieurs (utilisés en arithmétique, en algèbre, dans le calcul différentiel et intégral, en analyse mathématique, en logique mathématique). Pour cela, il fallait *concevoir des langages artificiels*, destinés à exprimer des algorithmes de calcul, *indépendamment des contraintes structurelles* imposées par les premiers ordinateurs et, en même temps, pouvoir *« traduire » automatiquement* ces langages artificiels dans les codes de la machine. En 1954-1957, J. Backus (considéré comme le « père » du FORTRAN) a conçu, avec son équipe chez I.B.M., la première vision de FORTRAN, premier langage de haut niveau. En même temps, il a réalisé le premier compilateur de cette version. Il a ainsi montré comment on pouvait concevoir, puis utiliser effectivement, des langages de « haut niveau » dont la structure était de plus en plus indépendante des implantations physiques sur telle ou telle machine informatique concrète. La méthode a été reprise, d'autres langages ont été proposés : pour les problèmes de calcul scientifique, le projet collectif d'ALGOL 60 reposait sur des bases théoriques plus solides que la première version de FORTRAN ; pour les calculs de gestion, le langage COBOL est issu de compromis entre les utilisateurs et des soucis de généralisations efficaces. Les informaticiens ont ensuite proposé de nouveaux langages de programmation, accompagnés de compilateurs ou d'interpréteurs : ALGOL 60 et ALGOL 68 qui ont servi de modèles à PASCAL (N. Wirth), APL (J. Iverson) ; LISP (J. MacCarthy) et PROLOG (A. Colmerauer)...

Ces langages de programmation restant indépendants des organes d'exécution, il fallait proposer une méthodologie générale d'écriture des programmes de transformation automatique du langage des utilisateurs dans les codes utilisables par des machines ayant des organes de traitement électroniques. Actuellement, on poursuit le mouvement de conception de nouveaux langages de différents types ; lorsque ces langages sont mis au point, on cherche ensuite à les implanter sur une machine réelle en construisant un interpréteur et/ou un compilateur, que l'on optimise progressivement. Les langages sont ainsi devenus mieux structurés avec une programmation par analyse descendante, ils permettent une manipulation de plusieurs sortes de données (ALGOL et PASCAL). Par exemple, un langage de programmation comme LISP sert à manipuler des listes (construction, fusion, accès aux éléments de la liste...) et à définir des nouvelles fonctions à partir de fonctions déjà définies ; ce langage permet d'écrire assez facilement des programmes pour résoudre des problèmes acceptant des formulations récursives. On envisage actuellement des langages de programmation plus « déclaratifs », c'est-à-dire des langages qui expriment des faits et ne s'occupent pas de « comment » ces faits seront manipulés, c'est donc au compilateur associé au langage déclaratif de déclencher, au moment opportun, les procédures de construction et de

recherche des informations nécessaires au calcul de résultats... Par rapport aux langages procéduraux qui sont centrés sur l'expression d'instructions impératives devant être exécutées dans un certain ordre par la machine, les langages déclaratifs autorisent un nouveau style de programmation, beaucoup plus « convivial » car beaucoup plus proche des habitudes de concevoir et d'écrire des utilisateurs humains.

Certains langages de programmation sont maintenant des superstructures d'autres langages. Le langage SMALLTALK, par exemple, est défini sur LISP. Un langage d'accueil pour des systèmes experts comme SNARK est défini en PASCAL. Un Langage Orienté Objet, comme MERING, a pour base le langage LE-LISP. Un programme étant écrit dans le langage LE-LISP, on peut traduire tous les fichiers de ce programme dans le langage C et faire ensuite appel au compilateur de C pour exécuter le programme initial. On peut également joindre les avantages de la « programmation logique » (d'un langage de programmation comme PROLOG) à ceux de la programmation fonctionnelle (c'est le cas de FUNLOG). Il y a ainsi toute une énorme variété de langages dont les structures ne dépendent plus uniquement des fonctions logiques booléennes. On sait maintenant « étendre » un langage en lui ajoutant d'autres fonctions écrites dans ce langage, on sait représenter les expressions d'un langage par des expressions d'un autre langage en ayant recours à des processus de compilation.

Pour mieux comprendre la compilation et en faire une théorie (non encore achevée lorsqu'elle atteint les représentations sémantiques), il a fallu approfondir les concepts de « langage », de « grammaire », de « syntaxe », de « traduction », d' « opérateur », de « séquence de symboles », de « symbole », notamment en regardant comment ces notions avaient déjà été appréhendées par les linguistes. A l'époque des premiers compilateurs, c'est-à-dire à la fin des années 1950, le linguiste N. Chomsky (1956) avait déjà entrepris ses premiers travaux en dégageant les principaux concepts utilisables dans une théorie de la compilation (langage formel, règles de réécriture, grammaire générative, automate d'analyse, ambiguïtés structurale et sémantique...). Ses études de linguistique formelle avaient été préparées par deux traditions, l'une linguistique, l'autre logique. D'un côté, des recherches descriptives en linguistique avaient été menées par des linguistes comme L. Bloomfield (1935) ou Z. S. Harris (1949), qui proposaient de véritables procédures algorithmiques obtenues en formalisant adéquatement les procédures distributionnelles. Harris, en particulier, aboutissait à dégager des structures quasi algébriques des énoncés des langues naturelles. D'un autre côté, des logiciens, comme A. Thüe (1914), E. Post (1943), A. A. Markov (1954), R. Carnap (1933), K. Gödel, A. M. Turing, W. V. O. Quine, Y. Bar-Hillel, avaient déjà proposé des définitions fécondes de

« alphabet », de « mot », de « système de relations et de congruences entre mots », de « langage », de « syntaxe », de « sémantique », de « traductibilité »... Les premiers concepts mathématiques d'une nouvelle branche des mathématiques, appelée depuis la théorie des langages formels et des automates, sont attribuables à N. Chomsky et à M. P. Schützenberger (1963). Ce dernier a proposé, en particulier, des méthodes mathématiques qui l'ont conduit à exhiber les propriétés algébriques profondes des langages formels (monoïdes ; langages rationnels et algébriques ; séries formelles...) qui permettaient de relier ces formalismes à l'algèbre non commutative. Ces travaux, essentiellement théoriques et mathématiques, ont certainement contribué à mieux fonder la théorie de la compilation des langages de programmation.

4. — PRINCIPE DE COMPILATION

Nous voulons généraliser la méthodologie employée en informatique et nous interroger sur son importation dans d'autres domaines, notamment dans les domaines du langage naturel. La compilation nous est apparue jusqu'à maintenant comme une méthode qui autorisait l'informaticien : (i) à concevoir (en anglais : « to design ») abstraitement et indépendamment des procédés d'exécution, des langages de programmation de « haut niveau » et (ii) ensuite, à écrire des interpréteurs et des compilateurs efficaces, chargés de « traduire » automatiquement les expressions et les programmes écrits dans les langages de haut niveau dans les codes des machines qui exécutent les instructions et les programmes. L'informatique a donc fait d'énormes progrès pratiques et théoriques dès lors qu'elle a su programmer la machine pour qu'elle « traduise » par elle-même et dans son propre code les algorithmes qui étaient présentés dans un langage algorithmique abstrait, proche des habitudes stylistiques des mathématiciens. En généralisant ce qui précède, on peut voir que l'informatique — principalement la théorie de la compilation — nous a appris que l'on pouvait :

(i) d'un côté, concevoir abstraitement des systèmes de représentations symboliques, plus ou moins déclaratifs et ayant des structures complexes qui tendent à être indépendantes des structures des organes et des supports physiques de la machine ;
(ii) d'un autre côté, ramener ces représentations à des représentations immédiatement compatibles avec les structures des organes et des supports (les V.L.S.I. dans nos ordinateurs actuels) ;

(iii) assurer le passage par un programme de compilation qui engendre des *représentations internes*, qui sont les intermédiaires entre les représentations externes, accessibles aux utilisateurs du langage, et les représentations digitales, accessibles aux machines.

La portée épistémologique et pratique de la compilation est considérable ; il est effectivement possible de manipuler des formes symboliques hautement structurées qui représentent des objets complexes (chaînes de caractères, tableaux, arbres, piles, listes, graphes, fenêtres, images...) en utilisant des machines dont les structures des composants de base sont différentes : il suffit de changer, par étapes successives, les représentations, de façon à se rapprocher du code des représentations supportées par les organes de la machine. C'est un des acquis les plus fondamentaux de l'informatique, que nous formulons sous forme d'un principe (que nous proposons d'appeler *« principe de compilation »*) :

Principe de compilation : pour manipuler du symbolique complexe par des composants naturels relativement élémentaires (c'est-à-dire physiques), une méthode éprouvée consiste à engendrer des représentations intermédiaires qui sont traduites les unes dans les autres par un processus de compilation.

La compilation est sans doute l'un des principes les plus fondamentaux de l'informatique. L'autre grand principe de l'informatique a été formulé par J. von Neumann — et peut-être, avant lui, par A. M. Turing : il s'agit de la notion de « programme enregistré » ; le programme est considéré pour la machine comme une donnée externe, il est mémorisé et peut, de ce fait, être appliqué à une série indéterminée de données toujours nouvelles. Les premières machines à calculer automatiques — celles de Pascal et de Leibniz — n'avaient pas de programmes enregistrés et ne permettaient donc pas d'y adjoindre des programmations de compilation au sens technique que nous avons rappelé.

En tenant compte du principe de compilation, et en dépit des commentaires médiatiques simplificateurs sur l'Intelligence Artificielle et l'informatique, il faut maintenant penser de façon beaucoup plus complexe les rapports entre les langages et les traitements sur des machines en renonçant :

(i) à établir une analogie entre les structures neuronales du cerveau et les structures des « commutateurs fonctionnant en 0/1 » ;
(ii) à considérer que l'ordinateur, objet artificiel, soit un modèle théorique, même approximatif, ou, comme il est souvent dit « une métaphore » du cerveau, objet biologique et naturel, ce qui reviendrait à vouloir affirmer une certaine analogie forte entre les fonctions du cerveau et celles de l'ordinateur ;

(iii) à « simuler » les comportements humains par des programmes informatiques qui ainsi seraient, sans modélisations théoriques et représentations intermédiaires, exécutables sur des machines à supports électroniques : les programmes donnant les mêmes résultats que les comportements observés, ils deviennent de ce fait instruments explicatifs des comportements humains.

Le programme de recherches qui s'appuierait sur les analogies précédentes est assez peu fécond car il est trop lié à un état du développement technologique ; de toute façon, il est peu soutenu par la communauté scientifique des chercheurs en Intelligence Artificielle. Il existe cependant, il est vrai, tout un courant philosophique qui reprend les vieux problèmes de l'esprit — *the Mind*— (opposé au cerveau — *the Brain*) ; ce dernier courant, défendu surtout par des philosophes, pénètre les sciences cognitives en proposant une certaine identification entre, d'une part, esprit et, d'autre part, les représentations des connaissances et les programmes informatiques : l'esprit serait au cerveau dans le même rapport que les logiciels, les structures des connaissances, les informations sont aux matériels, aux « puces » électroniques, aux supports physiques de nos machines à « traiter les informations ». L'analogie joue un très grand rôle dans ce type de débats. Ainsi :

« Comme on le fait souvent en sciences cognitives, il est utile de caractériser les fonctions des systèmes psychologiques par analogie avec l'organisation d'ordinateurs idéaux. [...] Plus précisément, si comme beaucoup de gens le pensent aujourd'hui, l'esprit est pour l'essentiel un système qui manipule des symboles, il devrait être intéressant d'étudier l'esprit par analogie avec les machines de Turing, car (toujours « en un certain sens ») les machines de Turing sont des systèmes de manipulation de symboles les plus généraux qui soient » (Fodor, 1986, p. 57).

Remarquons en passant que le terme de « computation » qui est souvent employé dans ce genre de débats n'est pas très clair et tend à obscurcir plutôt qu'à éclairer.

5. — COGNITION ET COMPILATION

Revenons maintenant aux problèmes centraux de la cognition : Comment établir des liaisons explicites et pertinentes, en se soumettant aux contraintes de la scientificité, entre le pôle du symbolique et celui du neuronal ? Comment les supports neuronaux traitent-ils effectivement les représentations symboliques ?

Plusieurs hypothèses de travail sont actuellement débattues. La notion de compilation reste souvent absente de ces débats ou est complètement marginalisée. Il en résulte, selon nous, des oppositions (comme symbolique/sub-symbolique) souvent peu claires, notamment lorsque l'on veut mettre dos à dos l'Intelligence Artificielle « orthodoxe » travaillant avec des représentations symboliques et les modèles connexionnistes travaillant avec réseaux d'automates interconnectés.

Certaines hypothèses (les hypothèses du connexionnisme et le paradigme du subsymbolique) soutiennent que les représentations symboliques, en particulier les représentations externes comme les représentations linguistiques, doivent *émerger* d'équilibres statistiques globaux obtenus par un réseau de composants simples. Certains travaux connexionnistes ont eu ainsi l'ambition de montrer comment le symbolique pouvait émerger statistiquement de réseaux neuronaux ; comment l'exécution des opérations devait être « hautement parallèle » et non séquentielle ; comment le codage des opérations et des informations devait être non localisable dans des adresses précises d'une mémoire mais devait au contraire être « largement distribué » sur les supports physiques. Certaines approches connexionnistes avaient visé au début à exclure tout recours à des symboles, la computation symbolique étant remplacée alors par des relations et opérations numériques, par exemple par des équations différentielles qui gouvernent un système dynamique ; le résultat d'une seule computation sur des symboles discrets serait obtenu à la suite d'un très grand nombre d'opérations numériques effectuées par un réseau d'automates simples. La signification ne serait pas donnée par les constituants eux-mêmes mais par des schémas d'activité complexes qui émergeraient d'une interaction entre certains d'entre eux.

D'autres hypothèses *(hypothèses des représentations intermédiaires et paradigme de la compilation)* insistent plutôt sur le caractère médiatisé des représentations symboliques et sur une certaine *architecture* nécessaire à toute organisation en différents niveaux de représentation et de contrôle. Entre les représentations externes et les représentations directement compatibles avec les modélisations effectuées par les neurosciences, il est postulé l'existence de représentations symboliques intermédiaires. Ces représentations intermédiaires ne sont ni des données, ni directement déductibles des données d'observations ; elles doivent être mises en place par un *processus abductif* prenant pour base les données empiriques effectivement observables et les analyses conceptuelles que nécessite toute entreprise de théorisation forte. Les niveaux de représentation (par exemple, dans l'analyse du langage : linguistiques, métalinguistiques, cognitifs, ..., neuronaux) sont reliés par des processus de compilation

(éventuellement paramétrés par l'environnement pragmatique). Ces niveaux ne sont pas isomorphes entre eux : chaque niveau acquiert une certaine autonomie et possède une structure qui lui est propre ; sa structure peut être différente de la structure des autres niveaux. Chaque expression du niveau i est le résultat de la compilation du niveau i-1, elle sert d'entrée au processus de compilation qui est orienté vers le niveau i + 1 ; la production de cette expression peut être *contrôlée* éventuellement par un autre niveau. L'ensemble des niveaux doit être organisé dans une architecture cognitive. Il ne s'agit évidemment pas d'une simple architecture hiérarchique qui établirait une chaîne de traitements séquentiels successifs (ascendante ou descendante), depuis les comportements externes jusqu'aux opérations effectuées sur un réseau de neurones.

Ainsi, on peut rechercher un niveau cognitif où langage, perception (des positions et mouvements spatio-temporels) et action (pointage dans l'espace, activités motrices) deviendraient compatibles ou échangeables. Pour cela, il faudrait que l'analyse de ces activités aboutisse à proposer un niveau de représentations où les représentations issues du langage et celles issues de la perception et de l'action motrice deviendraient, d'une part comparables et, d'autre part, seraient reliées aux observables linguistiques, perceptifs et moteurs par des processus de compilation différents. Si l'on est capable de mettre en place un tel *niveau cognitif de comptabilité,* on serait alors capable de montrer comment le langage ancre certaines de ses catégorisations grammaticales (relatives par exemple au temps et à l'espace) sur des catégories plus phénoménologiques.

On peut dire, en résumant, que l'hypothèse cognitive des représentations symboliques intermédiaires revient à adopter *la stratégie par compilation* :

> (i) en prenant acte des possibilités offertes par les processus informatiques de la compilation, au sens le plus large ;
> (ii) en construisant, par des procédures contrôlées par des abductions, des *représentations symboliques intermédiaires* prenant pour base des comportements humains observables, certaines de ces représentations ayant le statut de représentations mentales ;
> (iii) en se rapprochant progressivement, par un processus complexe de compilation, de représentations qui seraient compatibles avec les structures neuronales des supports biologiques (du cerveau) identifiables par les neurosciences ;
> (iv) en testant expérimentalement les opérations effectuées sur les représentations intermédiaires, y compris, bien entendu, les représentations mentales, par des implémentations exécutables sur des machines artificielles.

Les deux hypothèses (connexionnisme et représentations symboliques) ne nous paraissent absolument pas opposées, à condition toutefois,

comme nous l'avons dit, d'introduire dans le débat le *principe de compilation.* L'implémentation par compilation n'est évidemment pas liée aux seules machines séquentielles à mémoires adressables (machines du type von Neumann). On peut, en particulier, bien que ceci reste encore une recherche à faire, très bien compiler des représentations symboliques intermédiaires hautement structurées dans des représentations implémentables directement sur des réseaux formels de neurones. Les processus de compilation partent des représentations externes pour les transformer en différentes représentations symboliques de différents niveaux, en vue d'aboutir à des représentations neuronales formelles qui, d'un côté, seraient codées sur des supports physiques largement distribués et, d'un autre côté, conduiraient à des exécutions « hautement parallèles ».

Le principe de compilation est complémentaire de la notion d'émergence des niveaux subsymboliques (réseaux dynamiques, attracteurs, bifurcations, etc.) aux niveaux symboliques (symboles, règles...). Cette notion d'émergence du symbolique à partir du subsymbolique (P. Smolensky, 1986) a été bien commentée par J. Petitot (1989, p. 84-89) qui en a vu toutes les implications philosophiques et épistémologiques. Pour Petitot, il s'agit en fait beaucoup plus d'une « double émergence » et d'une « double organisation » (phéno-cognitive et phéno-physique) du monde naturel. L'émergence se doit d'expliquer « comment, à partir de configurations stables, le sens vient au symbole ». Par contre, la compilation a pour ambition principale de ramener des représentations hautement symboliques (manifestées par le langage, en particulier) à des implantations supportées par des organes physiques et biologiques, chargés d'effectuer toutes les opérations mises en jeu *par* et *sur* ces représentations. On voit qu'il s'agit là de deux conceptualisations qui ne s'opposent pas, mais doivent s'enrichir mutuellement.

L'analogie qui nous paraît exploitable et féconde dans les sciences cognitives est une *modélisation par compilation* et par représentations symboliques articulées à des implantations subsymboliques. Le cerveau, pour traiter des informations hautement structurées comme celles qui sont véhiculées par les langues naturelles et les raisonnements mathématiques, procède très certainement en mettant en jeu des compilations entre représentations, il engendre alors des représentations symboliques intermédiaires entre les expressions les plus externes et les expressions compilées. L'analogie, qui est pour nous féconde, doit être une *analogie entre des stratégies de traitement des informations* et surtout pas entre des organes ou même entre des fonctionnements.

Combien de niveaux intermédiaires entre les niveaux externes et les supports chargés d'effectuer les opérations ? Un seul niveau où toutes les activités cognitives seraient compatibles ou traductibles les unes dans les

autres ? Un « niveau conceptuel » unique, comme le prétend R. Jackendoff (1983), ou plusieurs ?

> « La thèse des fonctionnalistes est donc que pour l'étude de l'appareil cognitif, comme pour la description d'un ordinateur, on ne peut faire l'économie d'un niveau causal situé « au-dessus » du niveau physique. Dans les deux cas, il faut faire référence à des calculs effectués sur un système de représentations conformément à des lois abstraites. Le but est de mettre au jour un « langage de l'esprit » pour les opérations qui sous-tendent l'aptitude linguistique ou encore le raisonnement mathématique. [...] Le programme de recherche que proposent les fonctionnalistes serait de définir les différents niveaux pertinents de l'appareil cognitif avant de vouloir tout réduire à un niveau unique. [...] Peut-être, dans un siècle, en distinguera-t-on dix-huit ? » (Mehler, Dupoux, 1987).

Il s'agit de tout un programme de recherche pour au moins un siècle pour déterminer l'architecture cognitive adéquate.

Jean-Pierre DESCLÉS,
Université de Paris-Sorbonne,
Centre d'analyse et de mathématiques sociales,
Unité mixte C.N.R.S./E.H.E.S.S./Paris-Sorbonne.

BIBLIOGRAPHIE

ABELSON (Harold), SUSSMAN (Gerald J.), SUSSMAN (Julie), 1985, *Structure and Interpretation of Computer Programs*, Cambridge, MA, M.I.T. Press.

ANDLER (Daniel), 1986, « Les sciences de la cognition », in *La Philosophie des sciences aujourd'hui*, dir. Jean HAMBURGER, Paris, Gauthier-Villars.

BACKUS (John W), 1978, « Can Programming Be Liberated from the von Neumann Style ? A Functional Style and its Algebra of Programs », *Communications of the A.C.M.*, vol. 21, p. 613-641.

BLOOMFIELD (Leonard), 1933, *Language*, New York, Holt, Rinehart and Winston.

BONNET (Claude), HOC (Jean-Michel), TIBERGHIEN (Guy), dirs, 1986, *Psychologie, Intelligence Artificielle et automatique*, Bruxelles, Mardaga.

CHURCH (Alonzo), 1941, *The Calculi of Lambda Conversion*, Princeton, Princeton University Press.

CURRY (Haskell B.), FEYS (Robert), 1958, *Combinatory Logic*, vol. I, Amsterdam, North Holland.

CHOMSKY (Noam), 1956, « Three Models for the Description of Language », *I.R.E. Transactions on Information Theory*, IT 2, p. 113-124.

CHOMSKY (Noam), SCHÜTZENBERGER (Marc-Paul), 1963, « The Algebraic Theory of Context-Free Languages », *in* BRADFORD et HISCHBERG, *Computer Programming and Formal Systems*, Amsterdam, North-Holland.

DESCLÉS (Jean-Pierre), 1989, « Intermediate Representations in the Cognitive Sciences », *Semiotica*, 77-1/3, p. 121-135.

DESCLÉS (Jean-Pierre), 1990, *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès.

DREYFUS (Hubert L.), 1984, *Intelligence Artificielle : mythes et limites*, Paris, Flammarion.

FODOR (Jerry A.), 1986, *La Modularité de l'esprit : essai sur la psychologie des facultés*, trad. de l'américain Abel GERSCHENFELD, Paris, Minuit.

GANASCIA (Jean-Gabriel), 1990, *L'Âme-machine. Les enjeux de l'Intelligence Artificielle*, Paris, Seuil.

GLASSER (Huch), KANTING (Chris), TILL (David), 1984, *Principles of Functional Programming*, Englewood Cliffs, Prentice Hall International.

HARRIS (Zelling), 1951, *Methods in Structural Linguistics*, Chicago, University of Chicago Press.

HENDERSON (Peter), 1980, *Functional Programming : Application and Implementation*, Englewood Cliffs, NJ, Prentice Hall International.

HOFSTADTER (Douglas), 1986, *Bach, Gödel, Esher*, Paris, Interéditions.

JACKENDOFF (Ray), 1983, *Semantics and Cognition*, Cambridge, MA, M.I.T. Press.

MEHLER (Jacques), DUPOUX (Emmanuel), 1987, « La science cognitive », *Le Débat, 47*, nov.-déc. 1987.

PETITOT (Jean), 1989, « Hypothèse localiste, modèles morphodynamiques et théories cognitives : remarques sur une note de 1975 », *Semiotica*, 77, 1/3, p. 65-119.

PYLYSHYN (Zenon F.W.), 1984, *Computation and Cognition. Toward a Foundation for Cognitive Science*, Cambridge, MA, M.I.T. Press.

RICHARD (Jean-François), 1990, *Les Activités mentales. Comprendre, raisonner, trouver des solutions*, Paris, Armand Colin.

RUMELHART (David E.), MCCLELLAND (James), and the P.D.P. RESEARCH GROUP, 1986, *Parallel Distributed Processing*, vol. 1 et 2, Cambridge, MA, M.I.T. Press.

SMOLENSKY (Paul), 1986, « Information Processing in Dynamical Systems : Foundations of Harmony Theory », *in* RUMELHART *et al.*, p. 194-281.

STOY (Joseph E.), 1977, *Denotational Semantics : The Scott-Strachey Approach to Programming Languages Theory*, Cambridge, MA, M.I.T. Press.

SÉLECTION SÉMANTIQUE ET SÉLECTION NATURELLE
LE RÔLE CAUSAL DU LEXIQUE

Certains systèmes physiques — nous en l'occurrence — ont la curieuse propriété de pouvoir agir, et agir très spécifiquement, en fonction de ce qu'on leur dit. Il est parfaitement légitime, j'entends légitime du point de vue épistémologique, qu'on explique le pourquoi et le comment d'une action en remontant à un ordre verbal, ou écrit. En d'autres termes, un énoncé peut être la *cause* d'un certain comportement. Lorsque, selon notre psychologie spontanée, nous croyons bon d'expliquer une action *par* ce qui a été dit, nous présupposons déjà un certain déterminisme et une certaine finesse de grain de cette causalité. Malgré sa parfaite appartenance à l'ordre naturel, ce lien est ancré sur des structures abstraites hautement complexes et on n'a pas la moindre idée de la façon dont on peut l'analyser en des termes et des concepts propres aux sciences physiques. Dans cet article, je me propose de développer quelques considérations élémentaires sur la nature du lien causal entre actions et énoncés linguistiques, et en particulier sur le rôle fondamental que le lexique d'une langue naturelle joue dans cette causalité.

1. — LA CAUSALITÉ SÉMANTIQUE EST NATURELLE, MAIS PAS PHYSIQUE

A seule fin d'aider notre intuition, analysons, par exemple, les enchaînements d'actions spécifiques et distinctes, tous prévisibles, qui peuvent suivre ces différents ordres :
(1) Il faut le persuader qu'il doit partir.
(2) Il faut le persuader de partir.
(3) Il faut le faire partir.

Il nous est évident que (3), à l'opposé de (1) et (2), permet l'utilisation de menaces et chantages, tandis que (2) (un peu à la limite) et (3), mais pas (1), admettent la possibilité d'acheter carrément le départ par une

Revue de synthèse: IV[e] S. N[os] 1-2, janvier-juin 1990.

certaine somme. Il est aussi évident que cette différence s'explique en grande partie par la différence de signification entre les expressions « persuader » et « faire partir ». Ce lien d'une grande finesse entre actions et ordres, actions et suggestions, ou même actions et questions bien visées, fait partie de notre existence quotidienne. C'est bien sa « banalité » qui fait problème, du point de vue scientifique, puisque l'évidence même de ce lien démontre que nous sommes en présence d'un processus mental routinier, en grande partie inconscient, et, comme nous le verrons, de nature largement *déductive.*

Un autre aspect *central* de cette causalité est sa vaste indépendance par rapport au contexte. La spécificité des comportements qui peuvent suivre un ordre comme (1) ou (3) est pour nous généralement sans mystère, même quand on ne sait ni de qui il est question, ni quelles raisons motivent le locuteur. Il est clair, par exemple, en dehors de toute connaissance du contexte, que le locuteur de (1), (2) et (3) n'est *pas* la personne qui doit être persuadée, ou induite, à partir. Si tel était « l'état de choses » que le locuteur avait voulu exprimer ou qu'il avait voulu voir réalisé, il se serait plutôt exprimé ainsi :
(4) Il faut *me* persuader de partir.

La différence entre les actions qui suivent (4) et celles qui suivent (2) s'explique *entièrement* par la différence entre la contribution des mots « le » et « me » à la signification de chacun de ces énoncés. Il faut souligner aussi que la finesse de ce lien causal ne peut *pas* s'expliquer *autrement.* Cette causalité, *indépendante par rapport au contexte,* qui relie les énoncés aux actions, je l'appellerai, en m'inspirant des travaux de James T. Higginbotham (Higginbotham, 1985, 1986, 1988)*, une *causalité sémantique.* Causalité morpho-syntactico-sémantique serait peut-être un meilleur terme, mais il est trop lourd, et je préfère l'appeler simplement sémantique, après avoir dûment souligné au départ son indépendance par rapport au contexte. Je me propose, dans cet article, de montrer qu'il s'agit d'une causalité parfaitement naturelle. Mais il y a un certain intérêt épistémologique à souligner qu'elle ne se laisse pas réduire à une causalité *physique.*

En effet, les sciences physiques, telles que nous les concevons actuellement, n'ont aucun moyen de saisir les fondements mêmes de cette causalité. Par exemple, l'énergie acoustique véhiculée par les sons correspondants est totalement incapable d'expliquer le phénomène qui nous intéresse ici. L'univers des phénomènes sémantiques est étranger à l'activation sélective d'un système par l'intermédiaire d'une « touche » vocale.

* Pour plus de précisions concernant les références placées entre parenthèses dans cet article, se reporter à la Bibliographie, p. 90.

Il y a activation sélective, mais grâce à la médiation inéliminable de schémas abstraits, et non par l'effet d'une transduction directe des ondes de pression. A la différence de certaines machines dites « intelligentes » qui sont activées directement par la voix humaine, c'est-à-dire par les propriétés physiques de paquets d'ondes acoustiques, nos actions et nos pensées sont sélectivement engendrées par des structures abstraites que nous construisons *sur la base* des stimuli acoustiques. Nous *projetons activement sur* les paquets d'ondes acoustiques une famille de structures abstraites sous-jacentes, en correspondance étroite les unes avec les autres à plusieurs niveaux [1]. La causalité qui relie les expressions linguistiques à nos actions et à nos pensées *doit* faire appel à cette famille de structures abstraites et à leurs principes de correspondance ponctuelle, principes qui sont, comme nous le verrons par la suite, en partie universels et en partie spécifiques à une langue donnée. La structure physique du stimulus acoustique n'est certainement pas « simple », mais elle est néanmoins immensément plus pauvre que ce que notre appareil linguistique nous livre à la fin du processus de projection. Le résultat final du système de projections est bien ce qu'il nous faut pour fonder le processus de sélection sémantique. Parmi toutes les significations humainement concevables, une et une seule (parfois deux, et plus rarement encore trois ou quatre, dans le cas des expressions fortement ambiguës) sera choisie comme *la* signification de *cette* expression. Cette sélection est quelque chose que chaque être humain, dès son âge le plus tendre, fait normalement, sans effort, sans incertitude et presque instantanément. Il s'agit donc d'une sélection « naturelle », mais à ne pas confondre avec le processus évolutionniste darwinien qui porte le même nom. Les raisons de cette distinction apparaîtront plus clairement par la suite.

Il est important de souligner ici que les prétendues analogies que certains zoo-sémioticiens ont voulu établir entre le langage humain et la communication animale sont trompeuses. La nature abstraite, ouverte, infinie, compositionnelle, créative et générative des phénomènes sémantiques qui nous intéressent ici échappe à toute modélisation fondée sur un répertoire fini, discret et clos de commandes et de signalisations (voir les commentaires de plusieurs auteurs à Daniel C. Dennett, *in* Dennett, 1983). Comme l'a souligné Noam Chomsky (Chomsky, 1955, *in* Chomsky, 1985), les nombres naturels et les langues naturelles constituent les deux seuls systèmes « naturels » et biologiquement accessibles qui sont *à la fois* discrets et infinis. Aucune autre espèce animale ne peut

1. D'ordinaire ces niveaux reçoivent en grammaire générative les noms respectifs de : Forme Phonologique, Structure D, Structure de Surface et Forme Logique (CHOMSKY, 1981, 1986 ; MAY, 1985 ; RIEMSDIJK & WILLIAMS, 1986 ; LASNIK & URIAGEREKA, 1988).

se servir de systèmes possédant ces deux attributs à la fois, puisque toute communication animale est fondée, ou bien sur un répertoire discret et fini (de cris, de gestes, etc.), ou bien sur un canal à modulation continue d'intensité (phéromones, pirouettes, attouchements, etc.). La nature compositionnelle, abstraite, discrète et infinie de nos langues exclut que l'on puisse expliquer la sélection sémantique par une *liste* finie et figée de correspondances entre énoncés et significations. Il ne s'agit pas de trouver la touche intérieure (la signification) qui serait actionnée par la touche extérieure (le son) et qui, à son tour, déclencherait un comportement ou une représentation mentale. Il est fondamental de comprendre que la causalité sémantique ne se laisse reconduire à *aucun* schéma de ce *genre.* Même le lexique, comme nous le verrons par la suite, est un répertoire de structures abstraites de haute complexité, un ensemble fortement structuré et richement articulé, aussi bien au niveau global qu'au niveau de ses composants individuels. Un ensemble qui, de surcroît, demeure toujours « ouvert » à de nouvelles insertions (Richard J. Carter, 1984 a).

Rappelons aussi qu'il est impossible de saisir le phénomène de la causalité sémantique par des corrélations statistiques. Il est bien connu, par exemple, que les correspondances privilégiées (ou même *obligées*) à longue distance entre les termes d'un même énoncé défient le calcul des probabilités[2]. Qui plus est, la projection des structures syntaxiques semble engendrer obligatoirement des éléments « vides » *(empty categories)* qui sont cruciaux pour déterminer la signification, mais qui ne pourraient être soumis à aucun calcul statistique, pour la simple raison qu'ils ne sont pas présents du tout dans le stimulus manifeste (ils n'ont pas de « réalité phonologique ») et leur probabilité de parution est rigoureusement nulle (Chomsky, 1981, 1986 ; Lasnik & Uriagereka, 1988 ; Van Riemsdijk & Williams, 1986).

Un autre argument familier, qui souligne le manque d'intérêt des analyses statistiques en linguistique, est fondé sur un simple calcul combinatoire : n'importe quelle phrase de sept mots aurait une probabilité *a priori* ridiculement faible d'être présentée au locuteur. Plusieurs siècles peuvent tranquillement s'écouler sans que cet énoncé soit jamais présenté, même si une nouvelle phrase de sept mots était produite chaque seconde. D'où l'idéalisation courante en linguistique, selon

2. Il est *évident* à tous les locuteurs du français que dans la phrase : « Anne s'est demandé si c'était bien Jean qui a eu l'idée saugrenue d'interdire strictement à Ginette d'aller la voir à l'hôpital », il y a une forte présomption de co-référence entre « Anne » et « la », malgré le fait que ces deux mots sont séparés par 21 autres mots. Le calcul des probabilités ne parviendrait jamais à rendre compte de ces corrélations à longue distance (CHOMSKY, 1955 ; maintenant *in* CHOMSKY, 1985).

laquelle chaque énoncé a toujours la même probabilité *a priori* d'être présenté : la probabilité zéro (Chomsky, 1985 ; Elman, 1989 ; Marslen-Wilson & Tyler, 1980 ; Salasoo & Pisoni, 1985). Cette idéalisation élimine, entre autres, tout traitement de la sémantique basé sur la théorie classique de l'information, au sens de Shannon et Weaver. Il s'ensuivrait, en effet, qu'un énoncé quelconque, ayant une probabilité *a priori* égale à zéro, serait porteur d'une quantité d'information littéralement infinie. Il ne s'agit pas d'un paradoxe intéressant, mais d'une banale réduction à l'absurde, obtenue en appliquant un modèle très pauvre à un domaine qui n'est point de sa pertinence. La sémantique des langues naturelles est, tout simplement, étrangère à la théorie de l'information.

Souligner, comme je viens de le faire, que la causalité sémantique n'appartient pas au domaine des sciences physiques, ni à celui de la statistique, est parfaitement compatible avec la thèse que cette causalité appartient entièrement au monde naturel et que l'on peut, en principe, en trouver une explication scientifique. Sur la base de quelques contributions récentes (en partie encore non publiées), et sans prétendre en fournir ici une revue exhaustive, je voudrais présenter un schéma de solution naturaliste et non réductionniste qui me semble acceptable. La théorie que je peux à peine esquisser ici est une théorie naturaliste, compositionnelle et fortement innéiste des contenus sémantiques et des processus par lesquels ceux-ci peuvent être *la* cause aussi bien de comportements manifestes que d'états mentaux strictement « privés »[3]. En ce qui concerne le lexique, et donc les concepts « primitifs » (ou, plus exactement, les concepts qui sont mono-morphémiquement lexicalisables), ma théorie est atomiste, mais dans un sens un peu particulier : chaque signification lexicale est un *individu* doué d'une structure interne très riche et très spécifique. Chacun de ces atomes de la langue connecte *causalement* le locuteur de cette langue avec une certaine portion du monde, ou avec un certain état des choses, qui nous sont accessibles *sous une certaine description.* La causalité sémantique est une causalité *de dicto* et non pas seulement *de re.* Un point capital sur lequel nous reviendrons par la suite.

3. Il est sous-entendu que cette causalité présuppose toujours une certaine situation standard : que le message acoustique ou graphique ne soit pas trop perturbé par des facteurs de bruit, que le sujet soit physiquement, physiologiquement et psychologiquement capable d'entendre et d'agir, que ses compétences linguistiques et pragmatiques soient conformes aux normes courantes. Tout ce que je dirai ici fait tacitement appel à des conditions banales de *ceteris paribus* (FODOR, 1987, 1989).

2. — LA TENTATION BÉHAVIORISTE

Sous peine de circularité, une théorie causale naturaliste des contenus sémantiques doit agencer les éléments des langues naturelles (par exemple, le mot 'chat') avec des états de faits « publics » (l'animal de ce nom), sans *présupposer* déjà la nature foncièrement sémantique du lien. Une solution classique, celle de Skinner, consistait à dire que la *présence* physique d'un chat dans le voisinage du locuteur, sous certaines conditions normales d'illumination, d'attention, d'élaboration visuelle, etc., rend fort *probable* que ce locuteur prononce audiblement le son linguistique 'chat' (ou un son équivalent dans les diverses langues naturelles). L'idée centrale de cette approche bien connue est que la probabilité *a priori* d'entendre l'expression 'chat' est nettement plus grande lorsqu'il y a ou lorsqu'il y a eu quelques minutes auparavant, un chat aux alentours du locuteur. Une corrélation objective entre deux fréquences objectives (manifestations publiques d'objets et énonciations audibles de sons), tel est le fondement de cette sémantique béhavioriste.

Il devint vite clair, toutefois, que cette approche s'engouffrait dans des contrefactuels du type : si x avait été présent, au lieu d'y, alors on aurait observé X, au lieu de Y, avec probabilité Px. A bien y regarder, la « force » de ces raisonnements par contrefactuels se fondait toujours sur des relations nomologiques qui présupposaient la nature sémantique du lien et qui étaient, en conséquence, incapables de *fonder* ce lien. Un petit sommaire des déboires de la sémantique skinnerienne nous sera utile.

Il y a tout de même un petit point en faveur de la théorie béhavioriste, comme le souligne Jerry A. Fodor dans un manuscrit récent (Fodor, 1989). La théorie est, en effet, parfaitement « atomiste », car la connexion sémantique skinnerienne entre la bête et le son 'chat' pourrait tout aussi bien se matérialiser chez une espèce idéalisée qui ne possède, en tout et pour tout, qu'*un seul* concept : le concept 'chat'. En outre, même chez une espèce comme la nôtre, qui possède une énorme variété de concepts, le mécanisme causal qui relie Félix au concept 'chat' est, *de facto* et en principe, *indépendant* du mécanisme qui relie Médor au concept 'chien'. On a là deux mécanismes de *même nature* mais, dans l'univers des connexions causales entre concepts et objets, chacun de ces liens reste indépendant de tous les autres, aussi bien psychologiquement que, pour ainsi dire, métaphysiquement. Pour ceux qui, comme Fodor et moi, apprécient le caractère atomiste d'une théorie sémantique, cette indépen-

dance est un aspect positif. Malheureusement, bien d'autres aspects de la théorie skinnerienne ne le sont pas du tout.

Noam Chomsky a montré il y a longtemps (Chomsky, 1959) que la caractérisation de la situation « normale » de présentation de l'objet au locuteur se révèle être déjà foncièrement intentionnelle, c'est-à-dire, qu'elle présuppose tacitement la nature sémantique du procès qu'elle était censée expliquer. En bref, car l'histoire est trop bien connue pour qu'on s'y attarde, ce qui compte comme condition « normale » de présentation suivie par vocalisation ne se laisse nullement réduire à une liste « innocente » de situations physiques. Par exemple, le sujet prononcera le mot 'chat' si, et seulement si, il *pense* que quelqu'un aux alentours l'écoutera et sera *intéressé* à entendre l'expression 'chat'. Donc, la définition primaire de la normalité de la situation ne peut pas faire l'économie d'une composante mentaliste, qui est par nécessité bâtie sur des éléments qui sont *déjà* sémantiques. En outre, la théorie est totalement incapable d'établir une distinction entre les vocalisations (et donc les concepts) 'chat', 'chat-ou-lapin', 'félin', 'chat-avant-le-coucher-du-soleil', 'animal-qui-chasse-les-souris', 'animal-préféré-par-ma-grand-mère' et une infinité d'autres concepts dont la bête présentée ici et maintenant constitue *aussi* un exemplaire. N'importe quelle propriété qui co-varie causalement avec la présence de cet animal peut en fin de compte être, dans le schéma de Skinner, un candidat acceptable par la signification du mot 'chat'. Que dire, en outre, de la présentation d'un « stimulus » comme « la deuxième guerre mondiale » ?

Dans ce rappel sommaire des failles de la théorie béhavioriste, il faut au moins mentionner le problème de la disjonction et le problème des associations mentales (Dennett, 1987, 1988 ; Dretske, 1981, 1986, 1988 ; Fodor, 1987, 1989 ; Loar, 1981 ; Ruth Millikan, 1984, 1986, 1989 ; Segal & Sober, 1989). Un gros lapin que l'on entrevoit de loin par une nuit sans lune peut aussi causer la vocalisation 'chat', mais la signification de 'chat' n'est *pas* 'chat ou lapin-par-une-nuit-sans-lune'. La théorie sémantique béhavioriste n'a jamais pu résoudre ce problème redoutable et il y a de bonnes raisons de croire qu'elle ne *pourrait pas* le résoudre par ses propres moyens ultra-minimalistes et anti-mentalistes. L'autre problème insoluble pour la théorie béhavioriste est qu'une souris en plein jour peut aussi, par une association d'idées qui est parfaitement « normale », causer la vocalisation 'chat'. Comme Fodor le souligne, le problème de la disjonction et le problème des chaînes d'associations mentales guettent non seulement la théorie skinnerienne, mais toutes les approches naturalistes proposées jusqu'ici. Il nous faut introduire d'autres critères naturalistes, capables enfin d'expliquer *pourquoi* 'chat' est un concept primitif, tandis que 'chat ou lapin-par-une-nuit-sans-lune' est

un concept dérivé. En l'absence de contraintes *spécifiques* sur ce qui *peut* compter comme concept primitif, sur ce qui va être fixé *en premier* et de préférence lorsque nous sommes actuellement en présence d'un exemplaire de cette espèce, nous ne pourrions jamais décider si 'chat' veut dire 'chat' ou plutôt 'chat ou lapin-par-une-nuit-sans-lune'.

Un début d'explication nous vient des sciences cognitives (Carey, 1985 ; Keil, 1979 ; Markman, 1989 ; Smith & Medin, 1981 ; Pinker, 1989) et de la sémantique lexicale (Baker, 1988 ; Dowty, 1979 ; Fillmore, 1968 ; Fodor, 1987 ; Grimshaw, 1979 ; Hale & Keyser, 1989 ; Higginbotham, 1986 ; Jackendoff, 1983, 1987 ; Lederer, Gleitman & Gleitman, 1989 ; Levin & Tenny, 1988 ; Macnamara, 1982 ; Pinker, 1989 ; Pustejovsky, 1988 ; Rappaport & Levin, 1988 ; Talmy, 1985 ; Tanenhaus, Garnsey Boland, 1989 ; Tenny, 1988a, 1988b). L'acquisition des concepts par l'enfant est *contrainte* par une classe de concepts « naturels », qui sont toujours appris *en priorité* et sur la base desquels les autres concepts dérivés (notamment les concepts disjonctifs) sont *successivement* construits. Comme nous le verrons par la suite à propos de l'acquisition du lexique, il y a des significations (chat, par exemple) qui sont obligatoirement exprimées par un seul mot dans *n'importe quelle langue naturelle,* tandis que d'autres significations *doivent* être exprimées par des périphrases. Pour reprendre un exemple bien connu de Quine (Quine, 1960), mais avec des conclusions radicalement opposées à celles qu'il en tire, *aucune* langue humaine *possible* ne peut avoir un seul mot pour exprimer 'partie-de-lapin-non-détachée' et cinq mots pour exprimer 'lapin'[4].

Quine, qui se prétend skinnerien, croyait nous avoir appris à vivre paisiblement avec le problème de la disjonction, car il croyait nous avoir démontré qu'il n'y a, en principe, aucune façon de le trancher. Selon Quine, le choix entre 'chat', 'chat-ou-lapin', 'partie-de-chat-non-détachée', etc. reste à jamais indécidable, car il est dicté *uniquement* par des critères de simplicité pratique, par une convention tacite ayant une

4. Dans la fameuse situation imaginaire de Quine, un anthropologue-linguiste voit un lapin dans un pré et entend un indigène prononcer le mot *'gavagai'* en indiquant la bête. Quine soutient qu'il est impossible de décider en principe si la traduction correcte de l'expression *'gavagai'* est 'lapin' ou 'partie-de-lapin-non-détachée', ou 'lapinité-exemplifiée-ici-et-maintenant', ou une infinité d'autres expressions de ce type. Si *'gavagai'* est *un seul mot* de la langue de l'indigène, et si on n'a le choix qu'entre les trois traductions qui précèdent, alors, sans même y réfléchir, on sait que *'gavagai' doit* signifier 'lapin'. Cette connaissance est une connaissance innée, tacite et modulaire. Il n'y a aucune nécessité logique dans cela, mais il se trouve que notre système langagier est ainsi bâti. Les contraintes sur les significations possibles de nos lexiques naturels ne proviennent pas de la logique, mais de notre constitution biologique. Il se trouve que le choix entre 'lapin' et 'partie-de-lapin-non-détachée' comme candidats à la signification du son *'gavagai'* ne se pose même pas, ni à l'adulte, ni à l'enfant. Le problème de la disjonction est résolu au préalable par notre système langagier, c'est-à-dire, indirectement, par notre constitution spécifique.

certaine utilité sociale et par un principe de « charité » sémantique. Selon Quine, il est simplement sage et prudent de ne pas attribuer à autrui, sans y être contraint, un appareil sémantique trop différent du nôtre. Il s'agit là d'une bonne règle de parcimonie saine et utile, mais qui laisse le problème de la disjonction et de « l'indétermination de la traduction » entièrement ouvert.

La théorie sémantique de Quine est naturaliste, mais, à la différence de celle de Skinner, c'est une théorie globale (ou « holiste »), car le système des symboles joue comme un tout, chaque symbole recevant sa valence sémantique par sa position au sein du réseau de tous les autres termes de la langue. L'atomisme de Skinner est abandonné, en dépit d'une prétendue allégeance quinienne à la stricte doctrine béhavioriste. Ce que Quine nous offre c'est une théorie holiste et fortement indéterministe. Quine admet, toutefois, des contraintes biologiques, liées à l'évolution de notre espèce. En effet, ce qui unit plus profondément la sémantique de Quine et celle de Skinner, et les unit ensemble à beaucoup d'autres sémantiques naturalistes, c'est ce que j'appelle la tentation darwinienne. Encore une tentation à laquelle il faut résister.

3. — LA TENTATION DARWINIENNE EN SÉMANTIQUE

Tout semble se résoudre, même le problème de la disjonction, si l'on suppose que les « mauvaises » connexions sémantiques entre le monde et les significations ont été éliminées par les simples mécanismes de la survie. Depuis l'aube de ce qu'on appelle aujourd'hui l'épistémologie évolutionniste, c'est-à-dire depuis Charles Sanders Peirce (1896), Ludwig Boltzmann (1904), Konrad Lorenz (1941), jusqu'à nos jours, avec les théories sémantiques de Ruth Garreth Millikan (1984, 1986, 1989), Fred Dretske (1981, 1986, 1988), Daniel C. Dennett (1987, 1988), en passant par Karl R. Popper (1972), Donald T. Campbell (1974) et Willard Van Orman Quine (1974), cette solution a été retenue comme inévitable et parfaitement satisfaisante. Le lien offert par l'explication sélective darwinienne est bien un lien causal, car la démographie réelle des individus serait rigoureusement causalement covariante avec la démographie abstraite des contenus sémantiques. A la base de cette explication se trouve le raisonnement contrefactuel suivant : si nos mots et nos concepts avaient été mal « connectés » avec le monde, nous ne pourrions pas être ici pour en discuter, parce que nos ancêtres auraient été dévorés par les bêtes féroces ou anéantis dans des catastrophes naturelles. Puisque nous

sommes toujours là, il s'ensuit que notre sémantique « colle avec » le monde, du moins en première et bonne approximation. Ce que j'appelle ici la sélection sémantique serait, selon cette théorie, un effet *direct* de la sélection naturelle darwinienne.

Cette approche ne tombe pas dans la circularité de vouloir expliquer le sémantique par le sémantique, car, dans l'univers darwinien, on survit, ou on disparaît, objectivement, *sans même savoir pourquoi,* en fonction des concepts que l'on acquiert. Il ne peut, en outre, y avoir de lien plus « naturaliste », puisqu'il s'agit du *même* lien qui relie n'importe quelle structure biologique à ses fonctions, dans le cadre écologique « normal » de chaque espèce.

La théorie sémantique évolutionniste semble donc solide, causale, naturaliste et, si l'on attache de l'importance à ce point, aussi atomiste. En effet, les mésaventures sélectives par lesquelles les ancêtres des chats actuels ont fixé dans les esprits de nos ancêtres le concept 'chat' peuvent certainement être, en principe, causalement indépendantes des mésaventures par lesquelles les rochers de l'époque ont fini par fixer le concept 'rocher'. Comme il se doit, dans une théorie atomiste, deux mécanismes peuvent être de même nature sans pour autant dépendre causalement l'un de l'autre.

Les difficultés exorbitantes que rencontre *toute* théorie évolutionniste darwinienne des contenus mentaux ont fait l'objet d'un exposé détaillé (Piattelli-Palmarini, 1988b). Je me bornerai ici à en esquisser quelques aspects particulièrement pertinents pour mon propos.

Même en admettant que la théorie sémantique darwinienne puisse devenir aussi solide qu'il le faut, c'est-à-dire qu'elle puisse parvenir à corréler les objets du monde et les concepts d'une manière suffisamment forte et à l'épreuve des contrefactuels, elle ne sera jamais assez *fine,* elle n'arrivera jamais à être assez *spécifique.* Selon l'expression de Fodor, la causalité des mésaventures darwiniennes est insensible aux *descriptions sous lesquelles ces mésaventures nous sont arrivées* (ou plutôt, sont arrivées à nos ancêtres). En termes de survie différentielle, il est indifférent d'être dévoré par un tigre, par le seul félin qui possède des rayures sur fond jaune, par le seul mammifère qui possède la séquence génétique ...CCATTGG..., ou par l'animal que ma grand-mère déteste le plus. Mais les descriptions sous lesquelles les choses nous sont mentalement présentes, et sous lesquelles nous nous *rendons compte de ce qui nous arrive,* sont *essentielles* à la sémantique. Un être doué de raison, même d'une toute petite raison, ou d'une proto-raison, *doit* pouvoir se rendre compte de ce qui lui arrive et des raisons pour lesquelles cela lui arrive. Notre capacité de décrire, de présenter à autrui et de nous représenter mentalement, *que* x est un X, et non pas un Y, est bien indispensable à la

constitution même du tout premier embryon de sémantique. Pourtant, selon les mécanismes de la survie différentielle darwinienne, peu importe que la description de l'animal qui nous dévore soit X ou Y, et que la description sous laquelle *nous* sommes dévorés soit Z ou W. Ces mécanismes de survie darwinienne, bien connus (et utilisés de façon largement abusive), se moquent de la sémantique, parce que seul compte pour eux le nombre objectif des descendants et des disparus, en corrélation avec les causes génétiques et écologiques objectives. Le schéma darwinien n'arrive pas à engendrer une causalité *de dicto*, car il est *seulement* sensible à une causalité *de re*. Mais une causalité *de re* est *foncièrement* insuffisante à fonder une sémantique.

La théorie darwinienne ne peut pas sortir d'une double impasse en ce qui concerne la sémantique : si elle devient causalement dépendante des descriptions subjectives, elle perd sa force objective et elle présuppose déjà ce qu'elle était censée expliquer ; si, par contre, elle ne descend pas à ce niveau, elle reste à jamais étrangère à l'univers des significations. Ce qui donne à la théorie toute sa force (le compte objectif des cadavres et des survivants) lui soustrait du même coup la capacité d'engendrer une sémantique. Il me semble clair, donc, que les contraintes adaptatives dictées par la survie peuvent constituer, au mieux, une vague toile de fond sur laquelle la causalité sémantique reste encore entièrement à esquisser.

Les mécanismes sémantiques que nous possédons (comme bien d'autres structures et fonctions) sont, certes, *compatibles* avec un récit évolutionniste darwinien, mais ils restent radicalement sous-déterminés par celui-ci. Il y a une infinité de systèmes sémantiques possibles, qui seraient *tout aussi* compatibles avec les contraintes darwiniennes, mais qui ne sont pas les nôtres[5]. Seules les limites de notre imagination nous interdisent de concevoir en détail une *infinité* de systèmes sémantiques qui seraient

5. Par exemple, imaginons des êtres fort semblables à nous, mais chez lesquels tous les termes qui décrivent des objets solides font référence uniquement à la surface de ces objets. Doués d'un pareil système, nos ancêtres auraient parfaitement réussi à échapper aux tigres, à éviter de se cogner contre les rochers et à cueillir les fruits des arbres, mais il leur serait apparu « évident » que, par exemple, l'estomac du tigre n'est pas une partie de l'animal, mais quelque chose qui se trouve « près du tigre », tandis que les rayures font partie du tigre, car elles se trouvent sur la surface. Cette sémantique engendrerait, probablement, deux termes et deux « concepts » distincts pour la tortue (un lorsque la tête et les pattes sont dehors, un autre lorsqu'elles sont à l'intérieur). Des créatures raisonnables équipées avec un pareil système de représentations mentales se feraient, probablement, une idée un peu spéciale, que sais-je ?, des mystères du sexe et de la procréation (le fœtus serait, pour eux, 'près de la mère', et j'en passe). Pourtant, rien de cela ne rendrait ces êtres imaginaires, doués d'un système si radicalement différent du nôtre, incapables de survivre et de se multiplier. Tous les cas de significations authentiquement « impossibles » pour nous (voir, par exemple, les travaux de Richard J. Carter) seraient parfaitement compatibles avec la survie darwinienne.

pour nous tout à fait « innaturels », voire « inhumains », mais parfaitement compatibles avec les critères darwiniens de survie.

En sémantique, il faut donc *résister* à la tentation darwinienne, car il nous faut des contraintes naturelles immensément plus fines. Les contraintes qui peuvent *commencer* à fournir cette explication sont presque certainement de lointaine origine biologique, mais elles sont *radicalement sous*-déterminées par la simple survie différentielle darwinienne (Piattelli-Palmarini, 1989).

Ayant écarté malentendus et faux-départs, je me propose maintenant d'esquisser la nature des principales contraintes naturelles en sémantique.

4. — LES CONTRAINTES LEXICALES

Le caractère compositionnel de la sémantique de nos langues naturelles (Fodor & Pylyshyn, 1988 ; Higginbotham 1986 ; Carter, 1984 a, b) suggère que l'on commence par les unités élémentaires, c'est-à-dire par les termes individuels du lexique. Les contraintes naturelles sur les significations lexicales *possibles* seront très importantes pour nous. Sur ce point, toutefois, il nous reste encore à nous libérer de certains malentendus. La subtilité et la richesse des lexiques de nos langues nous ont joué un mauvais tour, car elles nous ont fait croire que n'importe quel concept, concret ou abstrait, simple ou complexe, individuel ou agrégé, ancien ou moderne, est *potentiellement* exprimable par un mot unique. Or une étude comparative de la structure universelle du lexique, en liaison étroite avec celle de la syntaxe et de la morphologie, a révélé de fortes contraintes sur ce qui est « nommable » (plus exactement, monomorphémiquement lexicalisable). Il ne s'agit nullement de contraintes agissant sur, ou dictées par, les limites de notre intelligence. Ces contraintes sont de nature exclusivement linguistique. Par exemple, il y a une infinité de choses que nous pouvons *aisément* exprimer par une circonlocution, mais que nous ne pouvons exprimer par un seul mot. Supposons que le verbe imaginaire 'fersuader' signifie « persuader une femme » et prenons un énoncé aussi simple que le suivant :

(1) Jean n'est pas capable de fersuader.

Nous ne savons pas si ce qui est nié par (1) est la capacité à persuader les femmes, ou à persuader tout court, ou les deux à la fois. Il nous est

difficile de décider ce que (1) *veut dire.* Le concept n'est certainement pas trop « difficile » pour notre entendement, pourtant nos stratégies naturelles, irréfléchies, de dérivation sémantique se trouvent bloquées. « Fersuader » n'est pas un concept mono-morphémiquement lexicalisable. Disons, par simplicité, que ce n'est pas un verbe *possible* d'une langue naturelle. Les choses deviendraient, en effet, très vite compliquées, si on considérait des énoncés tels que :

(2) Personne ne peut fersuader n'importe comment.

(3) N'importe qui peut être fersuadé par quelqu'un.

(4) Quelqu'un peut fersuader sans aucun effort.

Nous sommes sortis des limites naturelles d'élaboration et de sélection sémantique ; les procédés spontanés, irréfléchis qui nous guident sans aucune peine, instantanément, en présence des termes ordinaires de nos lexiques naturels ne peuvent plus nous aider[6].

D'autres exemples de mots (et de significations mono-lexicales) impossibles seraient : 'panger' (manger du pain et...) et 'blire' (lire la bible, mais pas...). Dès que l'on essaye des phrases qui nient, ou conditionnent hypothétiquement ces verbes imaginaires, ou opèrent sur eux avec des quantificateurs, on est perdu. La formation de simples mots composés, comme 'immangeable', 'illisible' (ou comme les équivalents de 'déboutonner', 'surcharger', etc.), se trouve aussi bloquée. Si l'on essaye ces constructions avec des verbes impossibles, on obtient des termes dont la signification nous échappe totalement. Par exemple, l'anglais admet par sa morphologie la combinaison de suffixes de comparaison et de superlativisation *(-er, -est)* avec des préfixes de négation *(un-, in-, counter-)* (comme dans *unhappier, unkindlier, unpleasanter,* etc.). La compréhension de la portée *(scope)* logique de ces modifications, pourtant si routinières, se trouverait totalement bafouée dans le cas de verbes impossibles (Pesetsky, 1985). Il nous faut réfléchir longuement avant de décider *ce qui a été dit.* Malgré ces efforts exceptionnels, il n'est pas sûr que nous puissions toujours y arriver[7]. Par contre, la négation et la quantification

6. Par exemple, quelle est la valeur de vérité de (2) s'il y a au moins un homme qui peut *p* ersuader n'importe comment n'importe qui ? Ou celle de (3) si, pour chaque femme, il y a toujours une autre femme, mais aucun homme, qui peut la *f* ersuader ? Que dire de (4), s'il se trouve que tous ceux qui peuvent persuader sans effort n'ont jamais essayé avec aucune femme ? Il nous faut quelques instants de réflexion avant de répondre, peut-être même un papier et un crayon.

7. Un exemple : « Aucun livre n'est assez banal pour ne pas être blu » (rappel : 'blire' = 'lire la bible, mais pas...'). Que faire d'une expression pareille ?

des termes des langues naturelles n'occasionnent jamais des difficultés de ce genre. Celles-ci montrent bien qu'il s'agit de contraintes spécifiquement lexicales, et non pas cognitives, ou computationnelles génériques. Il est, en effet, facile de trouver des périphrases qui peuvent exprimer aisément, et précisément, ces significations « impossibles », des périphrases que l'on découvre être d'une parfaite banalité conceptuelle.

Tout ceci témoigne que des termes imaginaires comme 'fersuader', 'blire' et 'panger' ne pourraient *jamais* devenir les termes d'une langue *naturelle* possible. Aucun enfant ne parviendrait à les « découvrir » dans une situation ordinaire (j'exclus ici la situation où l'on propose à l'enfant une devinette amusante).

Nous voyons donc qu'il y a des limites à ce qu'un seul mot *peut vouloir* dire dans une langue naturelle quelconque. Nous touchons là aux contraintes incontournables de notre système linguistique. Ces contraintes sont bien incontournables, quoique parfaitement *contingentes,* liées à notre biologie, et non pas à la logique pure, ou à l'efficacité de la communication. Pour les raisons analysées il y a un instant, ces contraintes ne proviennent pas non plus de mécanismes de la survie darwinienne.

Sur la base de considérations beaucoup plus techniques, que je ne peux résumer ici, il est apparu que ces contraintes sont parfaitement, quoique tacitement, accessibles à l'enfant lorsqu'il apprend sa langue maternelle. Une infinité d'hypothèses bizarres (comme 'fersuader', 'panger', 'blire', etc.), pourtant *toutes compatibles avec les données visuelles ou tactiles disponibles,* ne seront jamais envisagées par l'enfant. Non parce que l'expérience les contredit, mais parce que ce ne sont pas des hypothèses sémantiques humainement accessibles. Un extra-terrestre ou un ordinateur pourrait fort bien les envisager, mais pas un être humain. L'enfant ne supposera jamais que 'persuader' puisse faire référence *uniquement* à un des deux sexes[8]. Pourtant, il y a des termes du lexique qui se réfèrent à un seul sexe : sœur, maritorne, soubrette, fée, etc., mais aussi robe/costume, bas/chaussettes, etc. Cette subtilité du lexique, ainsi que la richesse des mots et des significations, défient le caractère, pourtant, si *évident* de ces multiples contraintes. Historiquement, cette difficulté nous a conduit à conclure qu'il n'y avait pas de contraintes du tout et que tout ce qui était pensable était aussi monolexiquement « effable ».

Il y a, au contraire, des contraintes sémantiques lexicales nettes et puissantes, qui sont pour nous si profondément évidentes et spontanées

8. Ou que 'chaussure' puisse signifier uniquement chaussure gauche, même si ce qu'on lui montre *est* une chaussure gauche.

que les linguistes arrivent à les dépasser seulement par un effort soutenu de l'imagination. Dans ce domaine, les généralisations scientifiques solides ne sont pas simples et les dernières années nous ont appris que la comparaison en profondeur de beaucoup de langues, couplée avec une analyse fine des effets syntaxiques de ces contraintes lexicales sont indispensables au progrès (Tenny, 1988 a).

Actuellement, les deux principales voies d'accès aux contraintes lexicales sont constituées par certaines conséquences sémantiques *déductibles* des contraintes morphologiques et syntaxiques, et par des cas hypothétiques de significations impossibles (tels que celui que je viens d'envisager) (Hale & Keyser, 1988, 1989 ; Jackendoff 1983, 1987 ; Pustejovsky, 1988 ; Tenny, 1988 a)[9]. La recherche sur la sémantique lexicale est en plein essor et il est difficile d'en anticiper les résultats. Je peux seulement en résumer quelques traits saillants. Il ne s'agit, pour l'instant, que de vérifier quelques-uns des *types* de contraintes agissant à notre insu sur le lexique de nos langues.

5. — UN CAS IMAGINAIRE : L'ACQUISITION DU VERBE « PLONCHER »

Le cas idéalisé que je veux développer ici est celui de l'enfant qui rencontre un mot, donc une signification lexicale, ou un 'concept', pour la première fois. Sur la base des travaux de James T. Higginbotham (1985, 1986, 1988), je veux reconstruire certaines contraintes indépendantes du contexte, et par là réfléchir sur l'enchevêtrement étroit de la morphologie, de la syntaxe et de la sémantique [10].

---

9. Une vraie floraison de significations lexicales impossibles a été brillamment engendrée et minutieusement examinée par Richard J. Carter au long des années (voir le recueil de ses écrits, paru en 1988 sous la direction de Beth Levin et Carol Tenny). Les exemples de Carter portent aussi sur la comparaison entre différentes langues et touchent à beaucoup de problèmes fondamentaux en morphologie, en syntaxe, en sémantique et aussi en philosophie du langage (un certain nombre de ces articles ont été publiés initialement en français ; voir, par exemple, CARTER, 1984c ; CARTER, 1980).

10. Des auteurs tels que Ray Jackendoff et Leonard Talmy font aussi intervenir des contraintes objectives, spatio-temporelles et « actantielles », sur les significations possibles. Certaines contraintes perceptuelles et conceptuelles générales, dictées par les structures topologiques de l'espace, par l'enchaînement des événements dans le temps, et par les types de mouvements possibles (ou concevables), seraient reflétées, selon eux, dans les significations lexicales. Cette approche a été formalisée mathématiquement (et aussi radicalisée quant à sa portée ontologique) par l'école « catastrophiste » française. René Thom et Jean Petitot tendent à donner à ces contraintes sur les significations un statut objectif formel, de nature intrinsèquement morphodynamique. Les structures sémantiques des langues naturelles

Je considère comme acquis que chaque énoncé manifeste est associé à une structure riche et spécifique, qui n'est que partiellement exprimée au niveau phonétique, mais qui est mentalement et tacitement représentée en tous ses détails à plusieurs niveaux d'abstraction (Chomsky, 1981 ; Lasnik & Uriagereka, 1988 ; Riemsdijk & Williams, 1986). Entre ces différentes représentations, il y a des correspondances ponctuelles obligées, qui ne sont pas accessibles consciemment au sujet (Chomsky, 1981, 1986 ; Lasnik & Uriagereka, 1988 ; May, 1985 ; Pesetsky, 1985, 1987 ; Riemsdijk & Williams, 1986). Ces structures cachées exercent des contraintes spécifiques et puissantes sur la signification du mot inconnu (ou imparfaitement connu, mais je veux ici analyser le cas le plus difficile). Pour rendre la situation intuitivement plus saisissante, je vais utiliser des mots inventés. Ce petit stratagème innocent nous permet de mieux nous « identifier » avec l'enfant.

Dans une situation que l'on peut aisément imaginer, l'enfant demande à sa maman ce qu'elle est en train de faire, et il entend la réponse suivante :

(1) Maman est en train de ploncher ce poisson.

Nous dirions que l'enfant n'a « aucune idée » de ce que ce verbe (imaginaire dans mon cas, mais réel dans la situation que je veux reconstruire) peut vouloir dire. Nous le dirions, mais nous avons tort, car la structure *syntaxique* de (1) véhicule *déjà* un bon nombre de contraintes sémantiques. Le fait qu'elles soient évidentes et banales n'empêche pas qu'elles soient capitales. « Ploncher » est quelque chose que maman (sujet) fait *au* poisson (objet)[11]. Il s'agit d'un verbe transitif (et qui exclut par le fait même d'innombrables significations comme dormir, neiger, flotter, évaporer, etc.). « En train de » signifie que cette action aura *par la suite* son résultat. Ce que maman a fait n'est pas *déjà* la « plonchade » ou le « plonchage », c'est une action partielle dans un schéma plus vaste, qui s'étale dans le temps.

Ces données syntaxiques excluent d'ores et déjà une quantité d'autres

---

seraient donc plus fondamentales, plus générales, plus profondes et plus révélatrices que leurs *conséquences* aux niveaux cognitifs, syntaxiques, etc. (PETITOT, 1989, 1985). Je me bornerai ici à souligner que l'extrême spécificité des données, sur lesquelles les linguistes de l'école générative travaillent, reste au-delà de la portée explicative de tels formalismes encore envisagés par Petitot.

11. Comme l'a montré en détail Lila Gleitman, même le tout jeune enfant déduit des significations différentes, spécifiques et parfaitement prévisibles, à partir d'un *même* dessin (ou photo), quand on lui décrit l'action d'un verbe imaginaire exemplifié par le dessin (ou la photo) comme : « A plonche B » par opposition à « A et B plonchent » (LEDERER, GLEITMAN & GLEITMAN, 1989).

significations impossibles. Ploncher ne peut pas être un verbe dit d'achèvement inéluctable, un de ces verbes qui caractérisent des actions telles que, lorsqu'on commence à en faire un tout petit peu, on les a *déjà* achevées (heurter, rencontrer, reconnaître, sursauter, etc.). Ce ne peut, non plus, être un verbe dit « vrai statif » *(true stative)* (avoir, savoir, aimer, ressembler, etc.), car aucun de ces verbes n'admet la construction « être en train de... » [12].

Il faut répéter que beaucoup de renseignements cruciaux sont *aussi* extraits par l'enfant de son observation visuelle, par exemple des gestes que sa mère effectue sous ses yeux, des outils dont elle se sert, etc. (il est crucial, par exemple, que ploncher est quelque chose que maman peut faire à elle toute seule, avec les outils disponibles dans la maison). Les sciences cognitives nous ont appris combien de contraintes perceptives, conceptuelles et sémantiques spontanées, non apprises, universelles, agissent aussi sur ces composantes perceptuelles, ontologiques et dans les attributions d'intentions à autrui (Carey, 1985 ; Gleitman, 1986 ; Goldstone, Gentner & Medin, 1989 ; Jackendoff, 1983 ; Keil, 1979 ; Landau & Gleitman, 1985 ; Markman, 1989 ; Smith & Medin, 1981 ; Spelke, 1985, 1988). Mais limitons-nous, pour l'instant, aux contraintes sur la signification qui se manifestent au niveau morpho-syntaxique.

Supposons que maman ajoute :

(2) Regarde chéri comme c'est facile de ploncher ce poisson.
(3) Ce poisson se plonche très facilement.
(4) Ce poisson est très facilement plonchable.

Une quantité d'informations nouvelles devient accessible à l'enfant. Pour nous en rendre compte, il suffit de comparer (4) avec des couples de verbes, pourtant presque synonymes, qui admettent, ou n'admettent pas, la construction-*able*. Par exemple : 'présenter-présentable', mais pas 'exhiber-*exhibable' ; 'naviguer-navigable', mais pas 'ramer-*ramable' ; 'surmonter-surmontable', mais pas 'excéder-*excédable' ; 'redouter-redoutable', mais pas 'craindre-*craignable' ; 'connaître-connaissable',

12. Il est intéressant de remarquer que cette construction peut *forcer* une expression à prendre parfois une signification dynamique, un devenir au lieu d'un être. Par exemple : « Il est en train de ressembler de plus en plus à son père » ; « Il est en train d'avoir de plus en plus d'argent ». En l'absence d'une expression qui souligne fortement la progression (comme « de plus en plus »), ces expressions seraient inacceptables (cf. *« Il est en train de savoir la réponse ») ou métaphoriques (« Il est en train d'aimer Marie »). (Je remercie Jean-Michel Roy et François Dell de m'avoir suggéré ces exemples).

mais pas 'savoir-*sachable'. L'absence de ces termes ne provient pas de considérations logiques, ou cognitives d'ordre général, mais semble être liée à des critères morphosyntaxiques extrêmement subtils [13].

Il est capital de souligner que, si ce dialogue s'était tenu en anglais, les informations auraient été encore plus riches et plus subtiles. Puisqu'il s'agit de contraintes sémantiques déduites (inconsciemment) de la syntaxe, il ne faut pas s'étonner que la traduction fidèle et exacte d'une phrase de la langue A dans une autre langue B puisse véhiculer *davantage* d'informations à l'enfant qui parle la langue B. En effet, si nous traduisons notre verbe imaginaire « ploncher » par le verbe anglais « to splonk », tout aussi imaginaire, nous obtenons l'équivalent de (3) :

(3a) *This fish splonks easily.*

Ceci dira implicitement, mais certainement, à l'enfant de langue anglaise qu'il ne s'agit pas d'un verbe de perception (comme *'see', 'hear', 'sniff'*, car on ne peut pas dire que quelque chose *« *sees easily »* ou *« *sniffs easily »).* En français, par contre, grâce à l'usage du « se », on peut dire que quelque chose « se voit facilement », ou « s'entend facilement » (Grimshaw, 1982 ; Kayne, 1975). Donc, en dépit de circonstances *identiques,* l'enfant de langue française qui entend (3) recevra des informations *différentes* sur ce verbe inconnu que l'enfant de langue anglaise qui entend (3a). Les deux « milieux » ou « situations » sont physiquement et psychologiquement identiques, mais *linguistiquement* différents. Et ceci est bien ce à quoi *il fallait* s'attendre, puisque les contraintes sémantiques qui nous intéressent ici sont de nature morphosyntaxique et variables d'une langue à l'autre (pour une analyse étendue voir aussi les travaux de Richard Carter, *in* Levin & Tenny, 1988).

Restons encore un instant dans le « milieu » syntaxique de l'anglais. (3a) exclut de nombreuses autres significations (Hale & Keyser, 1988, 1989) : les verbes instrumentaux sans modalité quantifiable en ce qui concerne leur objet (à la différence de ce qui concerne l'instrument lui-même), comme *'stab', 'nail', 'hammer',* et les verbes dits « vrais statifs », comme *'know', 'understand', 'trust',* car ceux-ci n'admettent pas la construction *.... *easily*[14]. Certains verbes, qui admettent cette construc-

13. Selon certains auteurs, il s'agit d'une structure lexicale projective, déterminée par le fait que le verbe peut ou non « c-commander » son argument interne, et par le nombre de niveaux de dérivation qui séparent la forme de surface et la forme logique (voir l'analyse présentée par PESETSKY, 1985).

14. Il suffit de penser à des expressions syntaxiquement inadmissibles comme : *'*This plank stabs easily', *'*This topic knows easily', *'*John trusts easily'* (dans le sens : « Il est facile de faire confiance à John »), etc.

tion (dite *middle construction* en anglais), admettent aussi une construction impersonnelle et sans spécification de l'agent. On peut dire *« This glass breaks easily »* et on peut, aussi, dire : *« This glass broke. »* Si l'enfant de langue anglaise entend une construction comme celle-ci *(« This fish splonked »),* il exclut d'emblée, *déductivement,* beaucoup d'autres significations intéressantes, car la mécanique de la situation pourrait ne pas les exclure par elle-même. Par exemple les actions, ou transformations, suivantes (je les transcris en français pour plus de simplicité) : 'saupoudrer', 'hisser', 'couper en tranches', etc.

D'autres contraintes sur la signification possible de 'ploncher' peuvent être tacitement exprimées par des phrases où ce verbe est enchâssé dans des expressions qui partialisent, ou quantifient *(measure-out)* l'action de 'ploncher' (Tenny, 1988 b). Par exemple, il y a une différence entre :

(5) Maman a plonché le poisson à moitié.
(6) Maman a plonché la moitié du poisson.
(7) Maman a mi-plonché le poisson.

On peut aisément, si tant est qu'on le puisse (c'est-à-dire, si la syntaxe du verbe 'ploncher' l'admet), construire d'autres phrases avec 'peu à peu', 'un peu à la fois', 'pendant une heure', 'en deux minutes', 'quasiment', 'presque', 'arrêter de...', etc. (Grimshaw, 1979 ; Hale & Keyser, 1989 ; Jackendoff, 1987 ; Pustejovsky, 1988 ; Rappaport & Levin, 1988 ; Tenny, 1988 a, b). Par exemple (5) signale (tacitement, je le répète) qu'il est possible de s'arrêter à mi-chemin dans cette action en ayant, du fait même, déjà fait une « moitié » de l'action, mais seulement une moitié. ('Ploncher' n'est donc pas comme 'frapper', 'secouer', ou 'trébucher', et bien d'autres verbes encore, car ceux-ci ne permettent pas ce genre de « partialisation » progressive). (6) signale qu'un autre type de partialisation est possible : celle qui consiste à compléter l'action, mais sur une fraction *d'un seul* objet seulement. Si on entend une phrase telle que (6), on peut *exclure* que 'ploncher' appartienne à la même catégorie que (parmi bien d'autres) 'poignarder', 'tuer', 'ranger', ou 'enfermer' (rappel : 'ce poisson' est un objet individuel, pas une collection d'exemplaires, et pas un amas d'éléments, comme le blé, la farine, le vin, le sable, etc.). (7) signale que l'action est, pour ainsi dire, progressivement mesurable à l'intérieur de son déroulement même, et que 'ploncher' n'appartient donc pas à la même catégorie que 'briser', 'interférer', 'livrer', etc.

Cette analyse pourrait bien s'étendre à d'autres phrases qui qualifient syntaxiquement l'action de 'ploncher' par des quantificateurs adverbiaux ('presque', 'quasiment', 'à nouveau', etc.), par des modalités psycholo-

giques ('intentionnellement', 'sans vouloir', 'avec tendresse', etc.), par des modalités effectuelles ('discrètement', 'en silence', 'gratuitement', etc.), ou par des temporalisations ('fréquemment', 'en deux minutes', 'pendant une heure', 'en une minute', etc.). La signification possible du verbe s'en trouverait progressivement de plus en plus circonscrite [15].

Il faut bien souligner que le rythme d'acquisition du lexique par l'enfant, surtout à l'âge de ce qu'on appelle (à juste titre) l'explosion lexicale, est prodigieux : en moyenne, un nouveau mot par heure, pendant des années (Miller, 1986 ; Miller & Gildea, 1987). Cette acquisition serait *impossible* s'il n'y avait pas de multiples contraintes morphosyntaxiques pour délimiter tacitement le nombre de « candidats » admissibles à la signification pour chaque mot nouveau que l'enfant rencontre. Dans un manuscrit en cours de publication, Jacques Mehler et ses collaborateurs montrent que des contraintes strictement phonologiques contribuent puissamment à signaler au tout jeune enfant le début et la fin d'un mot possible de sa langue maternelle. Ces contraintes phonologiques (par exemple, sur la structure des syllabes possibles et sur les positions possibles de l'accent tonique), couplées avec les contraintes morphologiques (Pesetsky, 1985) et les contraintes syntaxiques (Gleitman, 1986 ; Landau & Gleitman, 1985 ; Lederer, Gleitman & Gleitman, 1989), contribuent aussi à expliquer la rapidité et la finesse de l'acquisition du lexique. La mise en place du système sémantique naturel serait, en principe et en pratique, impossible sans l'aide tacite et puissante de ces principes morpho-syntactico-sémantiques, qui sont inconsciemment appliqués par l'enfant aux énoncés qui « entourent » la présentation initiale du mot, et aux termes dérivés à partir de ce mot.

Le peu que j'ai pu résumer ici de cette complexe procédure déductive et spontanée suffit déjà à nous montrer que ces principes et ces mini-théorèmes ne *peuvent* pas être eux-mêmes « appris » (Piattelli-Palmarini, 1989). Ils ne peuvent pas l'être, parce qu'on ne voit pas quel genre de données en fournirait à l'enfant des indices indirects univoques, et parce que personne ne saurait les « enseigner » explicitement à l'enfant. Ils restent encore largement inconnus aux linguistes de profession. Là où ils sont un peu venus à la surface, ils défient toute traduction en des termes que l'enfant pourrait comprendre dans un langage qui lui soit accessible [16]. Le procès semble être entièrement guidé par des ressources *internes*,

15. Par exemple, « en une heure » signale que nous sommes en présence d'une seule action, tandis que « pendant une heure » *(for an hour)* si c'est une action (et non un état) signale que nous sommes en présence d'une série de répétitions d'une même action.

16. Cf. note 13. Il est impensable que l'on puisse enseigner à l'enfant (et même à une majorité d'adultes) à maîtriser consciemment ce genre de critères.

probablement innées, inaccessibles à la conscience, et en grande partie communes à toutes les langues naturelles. Les spécificités syntaxiques des différentes langues (j'en ai fourni un petit exemple ci-haut) deviennent accessibles en vertu de principes morpho-syntactico-sémantiques *universaux* convenablement « paramétrisés » (Chomsky, 1981, 1986 ; Roeper & Williams, 1987). Dire que ces principes et ces contraintes pourraient provenir d'un enseignement implicite, indirect ou inconscient, reviendrait à poser un problème, et non pas à fournir une solution. En réalité, faire appel à un procès d'acquisition par « enseignement tacite », ou implicite ou indirect, équivaut à re-présenter le *même* problème sous une forme superficiellement différente.

Venons-en maintenant à la phase terminale du processus d'acquisition lexicale.

6. — L'ACQUISITION DE LA SIGNIFICATION « EXACTE »

Les contraintes que je viens d'esquisser font un bon travail, un travail indispensable, mais elles ne *suffisent* pas à déterminer de façon univoque ce que le verbe 'ploncher' veut dire exactement. Au terme de toute cette analyse tacite, l'enfant peut bien rester sur sa faim sémantique, ne sachant toujours pas si l'aspect crucial de cette action de 'ploncher' est à chercher dans une certaine action (écraser, farcir, saler, suspendre, fumer), ou dans un ensemble ordonné ou non d'actions. Une composante cruciale manque encore[17]. L'enfant n'est pas encore « connecté avec le monde » de façon sémantiquement correcte en ce qui concerne *ce* terme. Le manque de connection causale est, nous le voyons, atomique et non pas général.

La composante sémantique qui permet de sélectionner le candidat spécifique, la signification « exacte », est *foncièrement* constituée par une certaine suite d'actions et d'événements, *sous une certaine description.* Voici que les contraintes morpho-syntaxiques trouvent leur premier rôle *causal.* L'enfant est « connecté avec » un objet, une action, ou un événement, par une chaîne causale sémantique adéquate si, et seulement si, cet objet, cette action, cet événement, se présentent dans un certain encadre-

17. Lila R. Gleitman calcule que l'information morpho-syntaxique (ce qu'elle appelle le *syntactic bootstrapping*) fournit à elle seule une partie calculable de la signification d'un verbe (communication personnelle, voir aussi LEDERER, GLEITMAN & GLEITMAN, 1989).

ment préalable (*grosso modo,* par un encadrement du type de celui que nous venons de voir). Les données perceptuelles doivent toujours être autant de faits-sous-description. « Regarder » simplement ce qui se passe ne suffit pas. Il faut se « connecter » avec ce qu'on observe au moyen d'une grille de repères linguistiques. La causalité sémantique est, comme il se doit, une causalité *de dicto.*

Aux données perceptuelles et syntaxiques non sollicitées, s'ajouteront les réponses aux questions *pertinentes* de l'enfant, souvent basées sur des raisonnements par contrefactuels (« Maman, est-il possible de ploncher un poisson sur la plage ? », « Peut-on ploncher aussi de la viande ? », « Faut-il toujours avoir de la farine ? »). Comme il se doit, la causalité sémantique doit pouvoir soutenir des raisonnements hypothétiques bien choisis. Tel est le deuxième rôle causal des données syntaxiques : jouer finement avec des possibilités et des situations imaginaires.

Au terme du processus, la signification exacte du nouveau mot est fixée dans le lexique de l'enfant. Un nouveau concept a été acquis. Ce qui confère au processus son caractère « atomique », c'est la nature *extraordinairement* spécifique de chacun de ces liens. Comme nous le verrons, il s'agit d'atomes qui ont une structure interne, mais qui « correspondent » un par un à des significations exactes. L'aspect atomique du processus, que je tiens à souligner, n'exclut pas que l'on doive se servir, pour identifier un de ces atomes sémantiques, d'autres significations, d'autres mots et d'autres concepts. Pour décider de quel lien il s'agit, dans chaque cas spécifique, il faut examiner aussi d'autres liens et utiliser des comparaisons fines, en ayant recours à toutes les ressources de la morphologie et de la syntaxe. Mais faire appel à ce qui « entoure » le concept n'implique, ni ne présuppose que le lien une fois établi soit *causalement* dépendant des autres liens examinés. Des données holistiques peuvent bien servir à établir et à corroborer une hypothèse parfaitement atomique. L'important est que l'identité et la spécificité du lien ne dépendent pas *causalement* du « tout ». Elles peuvent en dépendre, toutefois, épistémiquement, de façon transitoire, pendant la phase d'acquisition.

7. — L'ORGANISATION INTERNE DES ÉLÉMENTS LEXICAUX

Les significations lexicales exactes ont deux composantes distinctes : une composante centrale et des informations associées, secondaires. Par exemple, il est essentiel à la signification de 'mariner' qu'il n'y ait pas de

cuisson, tout comme il est essentiel à la signification de 'persuader' qu'il n'y ait pas d'obligation [18]. S'il est indispensable, pour soutenir ceci, que l'on réintroduise la distinction entre analytique et synthétique, je veux bien être parmi ceux qui le font (Chomsky, 1986). Il s'agit, toutefois, d'une distinction un peu différente. Comme le mot l'indique, quelque chose est analytique (dans l'acception habituelle du terme), si on y a en principe accès par la réflexion, tandis que beaucoup de ces composantes sémantiques centrales ne le sont pas [19]. Dans mon schéma atomique, il y a une *structure interne* de chaque atome, des vérités ponctuelles qui ne peuvent pas être niées, sauf à changer *la* signification de ce mot (sans changer le concept). Les plus *importantes* de ces vérités, toutefois, ne sont que rarement accessibles à l'entendement par simple auto-clarification, elles doivent être découvertes par une analyse linguistique comparée, par des données expérimentales relevant de la psychologie du développement, de la psycho-linguistique, des neurosciences cognitives (par exemple, avec l'étude de cas pathologiques), etc. Les différences d'âge, de langue maternelle, d'éducation générale peuvent fournir des données sémantiques très importantes. La composante centrale de la signification de chaque mot est de nature proprement sémantique. Étant largement inaccessible à l'analyse par autoclarification, elle ne coïncide donc pas avec la composante analytique. Elle est aussi *sous*-déterminée par toutes les connexions causales objectives que le sujet peut observer dans le monde. La signification ne doit pas se surcharger de connexions causales objectives [20]. La composante centrale, en somme, n'est pas forcément bâtie

18. HALE & KEYSER (1988) et PUSTEJOVSKY (1988) soutiennent que la structure événementielle fine des significations conceptuelles lexicales est causalement responsable des « projections » syntaxiques fondamentales d'un verbe (nombre de rôles thématiques, caractère de ces rôles, possibilité de « décharger » ces rôles d'une certaine manière, etc.). HIGGINBOTHAM (1986) pense que seulement les rôles thématiques (ou *« theta roles »*) sont syntaxiquement importants, tandis que la structure fine des événements reste foncièrement « invisible » à la syntaxe. Tous ces auteurs, il me semble, se trouvent d'accord sur l'extrême spécificité et la grande richesse des structures lexicales internes.

19. Supposons, par exemple, que dans la signification du verbe 'voler' il y ait une composante inéliminable de volonté *soutenue* (volonté du sujet, ou de celui qui a projeté un engin, ou de celui qui le pilote). Une mouche, un hélicoptère et un avion en papier 'volent', mais pas un boulet de canon, ni une flèche, ni une feuille morte. Supposons que ceci soit vrai. Il me semble que cette composante sémantique centrale du verbe 'voler' n'est pas nécessairement accessible à l'introspection, ni dictée par un critère d'ordre logique. Aucune règle de l'entendement ne nous interdirait de donner au verbe 'voler' la signification 'se mouvoir dans l'air', sans autre précision, de telle manière que les pierres qui tombent, un boomerang et la fumée 'volent' aussi (un grand nombre de cas semblables ont été proposés par CARTER).

20. Pour aller à la chasse au renard il faut avoir des bonnes raisons de supposer qu'il y ait des renards dans la région, mais la *signification* du verbe 'chasser' ne se décompose pas en termes d'anticipations, de suppositions et de probabilité. Par contre, on ne peut pas dire que l'on 'chasse' quelque chose si on n'a pas l'intention de le capturer. Ceci est une composante cruciale de la signification du verbe 'chasser'.

seulement sur des « vérités de raison » et sur des « vérités de fait ». La composante analytique de la composante cruciale relève des principes de notre entendement, mais il y a aussi d'autres composantes, qui sont tacitement dictées par certaines structures modulaires contingentes de notre système cognitif (par exemple nos classes spontanées de similarité, nos prototypes naïfs et notre ontologie naïve). Elles ne demandent pas non plus à être vérifiées empiriquement par l'observation, car ce sont elles qui rendent l'observation *possible* en premier lieu. Leur fondement n'est pas dans les lois de la raison, comme c'est le cas des composantes strictement analytiques, mais dans notre nature humaine contingente.

Les vérités sémantiques centrales forment donc une classe particulière : leur composante analytique explique l'aspect *déductif* et les règles d'inférence logique mobilisées par le processus d'acquisition sémantique, tandis que la composante synthétique (largement tacite et ancrée sur des structures propres à notre espèce) explique l'aspect *inductif,* la recherche de confirmations et de réfutations spécifiques guidées par des hypothèses sémantiques spécifiques. La différence entre vérité analytique et vérité synthétique, que je tiens aussi à maintenir, ne *coïncide* donc pas avec la différence entre vérité sémantique et vérité d'ordre cognitif générique ou encyclopédique.

8. — SÉMANTIQUE LEXICALE ET CONNAISSANCE ENCYCLOPÉDIQUE

Je tiens à souligner trois points :

1) En temps réel et d'un point de vue subjectif, toutes ces composantes (morphologique, syntaxique, sémantique, cognitive, spontanée, inductive, encyclopédique) sont inextricablement entremêlées et leurs contributions respectives sont difficilement accessibles à l'introspection. Néanmoins, elles doivent en principe, et elles peuvent *de facto,* demeurer séparées dans une analyse scientifique. Il est clair, par exemple, que certaines composantes sont universelles, que d'autres sont spécifiques à une langue donnée, d'autres encore propres à une culture, ou à un certain groupe particulier, ou même à un individu donné. L'analyse scientifique parvient déjà à tracer des différences entre toutes ces contributions et à voir (ou entrevoir) les contraintes qui agissent sur chacune d'elles.

2) Les composantes morphologique, syntaxique et sémantique sont contraintes aussi bien par des principes de la grammaire universelle, commune à toutes les langues, que par des règles locales, propres à chaque langue (comme c'est le cas pour la *middle construction* en

anglais) ; la théorie linguistique actuelle postule un processus de « fixation de paramètres » pour rendre compte de cette relation entre universel et particulier.

3) La composante sémantique est enracinée dans le savoir encyclopédique de chaque sujet, mais elle est *en même temps* plus vaste et plus restreinte que ce que chacun « sait » du monde.

Sur ce dernier point, j'ajoute que le savoir sémantique individuel excède le savoir encyclopédique individuel, parce que chacun de nous est « fondé de pouvoir » en ce qui concerne l'usage *propre, plein et adéquat* de beaucoup de termes de notre langue (et même de beaucoup de termes de notre idiolecte privé), *en dépit* d'un accès seulement *partiel* à la signification exacte et à l'usage collectif attitré (par exemple, chez les scientifiques professionnels ou les experts), et en dépit de connaissances encyclopédiques souvent presque négligeables *sur ces mots* (Higginbotham, 1988 ; Piattelli-Palmarini, 1988a). Sous un angle fort différent de celui de Higginbotham et du mien, cette « division du travail » entre les experts et les locuteurs ordinaires fonde aussi la théorie causale de la référence de Saul Kripke (1972) et de Hilary Putnam (1975). Mais le savoir sémantique est *aussi* plus restreint, car seulement un sous-ensemble spécifique et limité de tout ce que nous savons sur les chats ou sur les objets volants est mobilisé dans la *signification* des mots 'chat' ou 'voler'[21]. En outre, il est crucial de souligner une fois de plus que le savoir sémantique exploite des éléments de connaissance tacite inaccessibles à l'introspection. Par exemple, comme nous l'avons vu ci-dessus, il est parfois sémantiquement capital que l'action exprimée par un verbe soit mesurable en ce qui concerne son « argument interne », mais pas son « argument externe », ou que la référence à un événement qui « culmine » soit obligatoire, ou que l'objet de l'action soit « affecté » d'une certaine façon, et ainsi de suite. Il s'agit de notions techniques, propres à la linguistique actuelle, qui échappent à la simple autoclarification introspective. Nous sommes censés avoir un accès introspectif à nos connaissances encyclopédiques, mais pas à ce genre de connaissances sémantiques. En effet, les connaissances sémantiques sont acquises, utilisées et mises à l'épreuve tacitement, sans trop de contrôle volontaire. Avec la plus grande facilité, un formidable appareil *individuel*

21. Par exemple, je considère qu'il *suffit* à un locuteur, pour pouvoir utiliser *correctement* dans son discours le terme « fractal », qu'il sache (en gros) qu'il s'agit d'entités mathématiques « découvertes » récemment, qui trouvent de plus en plus d'applications pratiques et que leur représentation graphique rappelle vaguement les vitraux d'une cathédrale. Par contre, un deuxième locuteur qui croit que les fractales sont une sorte de particules élémentaires, ne sait pas de quoi il parle, même s'il possède des renseignements encyclopédiques plus détaillés (et fondamentalement vrais) sur Benoît Mandelbrot, sur la date de leur « découverte », etc.

et de source interne sait comment se servir de données multiples (perceptuelles, linguistiques, déductives, inductives) pour parvenir à ses fins.

Il est utile de recourir, en sémantique, à maintes notions traditionnelles (par exemple à la distinction classique entre analytique et synthétique, à la notion de critère nécessaire et/ou suffisant, à la notion de fixation d'hypothèses, à la notion de définition, et ainsi de suite). Toutes sont utilisables en sémantique, mais aucune ne peut rendre compte, par ses seuls moyens, de la nature de la signification. Une meilleure compréhension des rapports entre savoir sémantique et savoir encyclopédique mobilise tous ces concepts, mais oblige aussi à cerner (rien de moins que) l'articulation entre les systèmes morpho-syntaxiques, perceptuels, cognitifs au sens large, et les systèmes de croyances, aussi bien collectives qu'individuelles. Les multiples tentatives pour réduire la sémantique à un et un seul de ces systèmes ont échoué. Ce qui n'a rien d'étonnant, vu la nature du problème.

Avant de réfléchir sur les conséquences philosophiques de ce schéma naturaliste, individualiste, internaliste et atomique, il reste à souligner que la signification « exacte » sera aussi exacte qu'il le faut, pas moins, et *pas plus*. Ceci est un point capital. Ma représentation mentale du mot 'mariner' n'est pas exactement la même que celle de Pierre. Néanmoins, nous entendons la même chose par ce mot. Nous avons la même représentation sémantique du mot 'mariner', en dépit de nombreuses différences entre nos images mentales individuelles (non seulement « en général », mais en ce qui concerne l'action spécifique de mariner). Aussi fine soit-elle, la connexion sémantique avec le monde n'impose ni ne présuppose *l'identité* psychologique des contenus mentaux individuels atomiques. Je laisse ici au niveau d'un simple énoncé ce qui pourrait être montré en détail : c'est-à-dire, que la causalité sémantique, selon mon schéma, est précisément celle qui permet *juste* le degré de variabilité qu'il nous faut parmi les représentations mentales individuelles soutenues (ou subsumées) par une *même* signification sémantique exacte. Les limites de cette variabilité coïncident avec le flou des informations collatérales que chacun possède sur un certain objet, ou sur une certaine action[22].

Ceci s'oppose, il est clair, à l'approche holiste de la sémantique, selon

22. Si Pierre croyait la cuisson *indispensable* pour 'mariner', ce mot aurait, dans l'idiolecte de Pierre, une autre *signification* que dans mon idiolecte. Si, par contre, Pierre croyait que pour 'mariner' il est indispensable de tremper quelque chose dans le vinaigre (pas dans du citron, ou de la sauce, etc.), il y aurait, entre Pierre et moi, seulement une différence d'ordre encyclopédique. Nous aurions la même sémantique pour le mot 'mariner' (nous en aurions le même concept), en dépit de la différence de nos croyances sur le *processus* (pour une analyse exhaustive de ces cas, sous une perspective différente de la mienne, voir les travaux de Burge).

laquelle *toutes* les informations collatérales (toutes nos connaissances encyclopédiques) sont *en principe* inséparables de la signification proprement dite. Les holistes soutiennent qu'il n'y a aucune distinction de principe entre la signification *(meaning)* d'un mot et les croyances *(beliefs)* portant sur l'état de choses que ce mot symbolise. Selon eux, une signification « n'est rien d'autre que » le faisceau variable des croyances que chaque locuteur entretient et qui sont *vaguement* véhiculées par ce mot. Il est typique de cette approche de nier, en principe également, toute séparation entre syntaxe, sémantique et pragmatique et de nier avec force la séparation entre synthétique et analytique[23]. Pour un vrai holiste, il n'y a jamais *la* signification « exacte » d'un terme isolé. Il n'y a que des réseaux de significations se soutenant les unes les autres et s'interdéfinissant par coopération.

Cette extrême libéralité sémantique n'a pas satisfait tout le monde et on a cherché à y mettre quelques limites. Le projet devint alors non pas de réintroduire une distinction entre le synthétique et l'analytique, mais de régimenter l'usage des concepts et des mots par des critères robustes de cohérence inférentielle subjective (Tyler Burge), ou par des chaînes causales objectives, largement indépendantes des contenus mentaux du locuteur particulier (Putnam et Kripke). Il faut admettre que la démarcation entre composantes cruciales et composantes secondaires de la signification a connu des difficultés considérables. Je ne peux pas les passer sous silence.

9. — OBJECTIONS ET CONTRE-OBJECTIONS

L'espoir de dégager un noyau central fixe dans la signification d'un mot a été fortement entamé par la démonstration de Putnam selon laquelle il n'est même pas indispensable à la signification du mot 'chat' de croire qu'il s'agit d'un animal (et pas, par exemple, d'un robot), et par celle de Fodor selon laquelle 'tuer' ne peut pas être dérivé sémantiquement de 'causer la mort'. En sémantique, Putnam et Kripke ont montré l'insuffisance foncière des critères nécessaires et suffisants (Kripke, 1972 ; Putnam, 1970, 1975), tandis que Fodor a montré l'insuffisance foncière des définitions (Fodor, 1970, 1987). Ces auteurs en ont tiré des conclusions opposées : pour Putnam et Kripke, les significations forment un

23. Quant à la morphologie et à ses effets sémantiques indéniables, il me semble que les holistes n'y aient même jamais pensé.

réseau en transformation continue, engendré par une interaction causale *collective* et *objective* avec le monde ; pour Fodor, les significations sont atomiques, engendrées (une par une) par une interaction causale, bien déterminée, individuelle et strictement atomiste. A la différence de ce que je soutiens ici, Fodor pense qu'il n'y a *pas* de structure lexicale *interne*, et que chaque mot, chaque concept, est comme une ampoule indécomposable, séparée des autres ampoules. Elle s'allume dans notre tête si, et seulement si, nous sommes connectés causalement de la bonne manière avec une certaine région (ou un certain aspect) spécifique du monde[24]. La sémantique est, pour Fodor, une propriété intrinsèque du « mentalais » *(the language of thought)* qui se *reflète* seulement de manière indirecte dans les traductions standard dans les différentes langues naturelles et constitue les contraintes objectives sur la syntaxe de ces traductions. Ce que j'ai appelé ici sémantique serait, pour Fodor, de la *pure* syntaxe des langues naturelles, sous un autre nom.

Certaines de ces objections me semblent insoutenables, certaines autres justes, mais contournables. L'hypothèse d'une construction foncièrement collective et holistique de la sémantique me semble réfutée par la découverte d'un nombre croissant de ressemblances (et identités) morpho-syntactico-sémantiques fondamentales entre toutes les langues, y compris les langues historiquement et géographiquement éloignées. Il faut en conclure que la prétendue chaîne causale collective est inexistante, ou interrompue depuis fort longtemps, et qu'on ne *peut* pas expliquer ces ressemblances profondes, entre des langues objectivement isolées l'une de l'autre, par une transmission collective de bouche à oreille. Qui plus est, comme je l'ai indiqué, la nature même des principes morpho-syntactico-sémantiques fondamentaux est telle que leur accessibilité à la conscience est pratiquement nulle. La transmission ne peut pas se fonder sur un enseignement explicite. Invoquer la possibilité d'un enseignement implicite (même là où elle est concevable) n'est qu'une autre façon de poser le problème. L'efficacité et la vitesse du processus d'acquisition lexicale peuvent s'expliquer *seulement* par une utilisation massive par

24. Il n'y a pas d'électrons et de protons dans les atomes sémantiques de Fodor. La connexion causale qui fonde la sémantique à la Fodor n'est pas quelque chose à laquelle on doit chercher un « accès » psychologique, ou linguistique, quelconque. C'est simplement une relation objective qui nous implique, quelque chose qui nous arrive, et non pas une relation que nous engendrons par nos opérations mentales. Celles-ci se mobilisent par un tout autre mécanisme, celui qui engendre nos attitudes mentales (croire, redouter, espérer, nier, affirmer, etc.) vis-à-vis de nos concepts. Tout ce que j'ai dit plus haut sur le lexique et son acquisition est, pour Fodor, un aspect de la *traduction* de nos concepts du 'mentalais' dans une langue naturelle (anglais, chinois, etc.). Comme il l'admet lui-même, certains aspects de sa conception actuelle du lien causal objectif entre le monde et les atomes sémantiques du mentalais sont (curieusement) assez proches d'une conception skinnerienne (FODOR, 1989).

chaque locuteur de contraintes non apprises, indépendantes du contexte et de nature strictement tacite.

A la différence de Fodor, je suis toutefois persuadé que les atomes sémantiques ont une structure interne et que certaines composantes sont cruciales pour la signification. Si un locuteur n'a pas compris qu'elles sont cruciales, il n'a pas vraiment acquis la signification. Par contre, d'autres composantes sont accessoires et pour elles j'admets une large variabilité de croyances et d'usages entre des locuteurs qui attachent, néanmoins, la *même* signification aux mêmes mots.

Un problème de redondance se manifeste d'emblée, puisque la frontière entre ce qui est crucial et ce qui est accessoire *dépend* du terme en question. A la limite, il se peut bien qu'il y ait autant de composantes cruciales qu'il y a de concepts mono-lexicalisables. Je considère toutefois le problème de la redondance comme un pseudo-problème, car ici, comme partout ailleurs en biologie, il ne faut pas présupposer de parcimonie (Piattelli-Palmarini, 1985), ni chercher à tout prix une économie de moyens. Le rôle causal des composantes sémantiques cruciales n'est *pas* celui de minimiser, ou de réduire, la tâche d'acquisition, le poids de mémorisation, ou la complexité d'usage. Leur rôle est de fonder l'invariance de la signification et la spécificité de la causalité sémantique, en dépit des différences cognitives individuelles.

Pourtant Putnam a raison : les composantes cruciales ne sont *pas* identiques aux critères nécessaires et suffisants de nature encyclopédique, structurale ou micro-physique. Fort peu de locuteurs savent ce qui est biologiquement nécessaire et suffisant pour décider qu'un animal est un chat, physiquement nécessaire et suffisant pour décider que quelque chose est en aluminium, pragmatiquement nécessaire et suffisant pour décider quand il y a eu une 'interférence'. Tous les locuteurs, pourtant, savent ce que *signifient* 'chat', 'aluminium', 'interférence'. Ce qui semble crucial, pour saisir la signification de ces termes, c'est plutôt (en première approximation) la propriété de miauler et d'avoir des moustaches, la brillance et la légèreté du matériel, l'entrave à une action entamée et conduite par autrui. Je ne me préoccupe pas de mieux « définir » ces composantes, car il ne s'agit *pas* non plus, comme Fodor l'a justement souligné, de *définitions*. Une définition utilise une périphrase bien polie, constituée par des mots qui nous sont tous déjà explicitement connus. Les composantes sémantiques cruciales, par contre, peuvent nous être connues seulement de façon tacite, sans qu'elles soient « effables » (tandis qu'une définition *doit* l'être) et sans qu'elles soient accessibles à l'introspection (tandis qu'un critère nécessaire et/ou suffisant *doit* l'être). Ces composantes font l'objet d'une véritable recherche scientifique empirique, guidée par des théories morphologiques et syntaxiques profondes

et détaillées et, le plus souvent, par une comparaison avec des langues fort « exotiques ». Ces composantes cruciales se révèlent parfois fort différentes de ce que notre intuition nous suggère[25]. La sémantique des langues naturelles mobilise tacitement aussi toutes sortes de « théories » naïves sur le monde, qui sont bien les nôtres, mais sur lesquelles nous exerçons peu de contrôle conscient (Carey, 1985 ; Jackendoff, 1983 ; Keil, 1979 ; Macnamara, 1982 ; Markman, 1989 ; McCloskey, 1983 ; Nersessian & Resnick, 1989 ; Spelke, 1988).

10. — LES COMPOSANTES CRUCIALES DES ATOMES SÉMANTIQUES

L'acquisition portera, non pas pour des raisons logiques ou pour des nécessités d'ordre pragmatique, mais pour des raisons contingentes à notre espèce, d'abord sur les ensembles naturels typiques *(natural kinds)* et sur leurs attributs fondamentaux, sur les actions fondamentales et sur les types d'événements fondamentaux. Ces catégories ontologiques spontanées correspondent, dans l'ordre, aux catégories syntaxiques et morphologiques de syntagme nominal ('ta sœur', 'les bébés', 'un ours', etc.), de syntagme verbal ('s'en va', 'est arrivé', 'disons', etc.) et syntagme adverbial-modal ('lentement', 'hier', 'sans bruit', etc.). L'ontologie spontanée pré-sélectionne les unités saillantes et leurs attributs (manifestes ou cachés)[26]. La psychologie spontanée sélectionne les actions saillantes et permet d'en deviner les causes internes, volontaires ou involontaires. L'usage de plus en plus subtil des quantificateurs syntaxiques ('le', 'un', 'plusieurs', etc.), des quantificateurs logiques ('tous', 'aucun', 'un seul', 'quelques', etc.) et des quantificateurs adver-

25. Par exemple, en dépit de ce que pensent beaucoup de locuteurs de l'anglais, ce qui est crucial à la signification du terme *crunchy* n'est point une certaine consistance matérielle, mais un certain *son* crépitant que l'on *doit* entendre dans la bouche. Les locuteurs du français semblent être mieux situés pour capturer intuitivement cette composante sémantique du mot correspondant 'croquant'. Pourtant, les deux termes sont la traduction fidèle l'un de l'autre. Ces différences entre les langues sont souvent profondes et très révélatrices, mais il serait insensé d'en tirer des arguments pour une quelconque intraduisibilité systématique entre les langues naturelles. Une meilleure compréhension de la sémantique permet, entre autre, d'expliquer les raisons du phénomène opposé : il y a traduisibilité systématique entre n'importe quelle langue naturelle et n'importe quelle autre langue naturelle. Les cas d'authentique intraduisibilité sont exceptionnels, quoique fort intéressants pour le théoricien.

26. Déjà à l'âge de deux ans et demi un enfant sait parfaitement quand un terme fait référence à un objet (parapluie, marteau) et quand un terme fait référence à un agrégat, ou une substance, (pommade, sable, etc.) (travaux de Susan Carey et Nancy Soja).

biaux ('souvent', 'jamais', 'avant de', 'pendant que', etc.) permet à l'enfant de saisir, et de proposer à ceux qui l'écoutent, des situations réelles *et hypothétiques* de plus en plus fines. D'autres termes et d'autres significations deviennent accessibles (conditionnels, conjonctions, disjonctions, relativisations, etc.). Le *bootstrapping* syntaxique permet d'aiguiser les nouvelles acquisitions lexicales jusqu'à un degré de raffinement comparable à celui de l'adulte. L'enfant sait tacitement, sans que personne ne le lui ait jamais « enseigné », que seules certaines connexions sont cruciales et qu'il lui faut, cas par cas, les détecter. Les ressources internes, aussi bien cognitives que linguistiques, sélectionnent ce qui *peut* constituer une composante cruciale de *cette* signification lexicale[27]. Comme le souligne Higginbotham (Higginbotham, 1988), il peut y avoir un élément *indiciel (indexical)* irréductible dans le mécanisme causal sémantique. On *pointe* vers un terme ou une expression de la langue *et* vers quelque chose de spécifique dans le monde, et on laisse les ressources internes compléter la définition de ce vers quoi l'on pointe, le degré de détails, la finalité de l'opération, etc. Le questionnement pertinent est une méthode fondamentale pour affiner le lien causal entre le sujet et ce qui vient d'être montré. On sait bien qu'il n'y a point d'ostension sans conceptualisation et les ressources internes fournissent *juste* la conceptualisation nécessaire[28], pour que l'acte d'ostension soit un acte sémantiquement réussi. Ni plus, ni moins.

L'idée centrale de mon schéma est qu'une fois détectées, les composantes cruciales d'un concept vont constituer *le* lien causal cherché entre *la* signification et *cet* 'état-de-choses-sous-description'. Le lien est ponctuel, spécifique et capable d'engendrer par *déduction* tacite toutes sortes de conséquences spécifiques, à plusieurs niveaux (morphologique, syntaxique, sémantique et cognitif). La composante syntaxique se charge, bien sûr, d'agencer les mots en énoncés de plus en plus longs et complexes, doués de significations de plus en plus subtiles. On parvient à décrire des états de choses qui sont vrais ou faux, ou probablement

27. L'intention du constructeur et l'usage habituel prévu seront des bons candidats pour un terme qui fait référence à un produit industriel ou artisanal, mais pas pour le nom d'un animal, tandis que le changement effectué dans l'état d'intégrité matérielle d'un objet sera un bon candidat pour un verbe transitif de manipulation, mais pas pour un verbe intransitif faisant référence au temps atmosphérique. Et ainsi de suite.

28. C'est par des ressources internes que chacun sait que l'adjectif 'blanc', appliqué à un œuf ou à un mur, n'implique pas qu'ils soient blancs de part en part, tandis que l'adjectif 'jaune' appliqué à l'or implique que cette substance soit jaune de part en part. Il serait impossible de connecter causalement l'enfant au monde en lui « montrant » quoi que ce soit, s'il était dénué de ces hypothèses restrictives spontanées. De toute évidence, ces choses-là personne ne nous les a « enseignées », car les enseigner présupposerait toujours que des connaissances *tout aussi complexes* que celle-ci soient déjà accessibles.

vrais, ou probablement faux, à partir d'atomes qui sont « vrais de » *(true of)* certains objets ou de certains événements atomiques.

Le locuteur, en somme, possède tous les outils qui lui sont nécessaires et suffisants pour « connecter » un certain énoncé avec un certain état de choses. En vertu de sa nature humaine et de la spécificité linguistique de son environnement (qu'il doit « apprendre » en même temps qu'il enrichit son lexique), il se trouve *causalement* connecté avec le monde, et avec ses interlocuteurs, par des mécanismes efficaces, subtils et hautement sélectifs. La causalité dont je parlais au départ, celle qui explique nos actions spécifiques par des expressions spécifiques de nos langues naturelles, est enracinée dans les mécanismes que je viens d'esquisser. Elle est naturelle, car elle est ancrée dans notre patrimoine biologique, mais elle n'est pas réductible à la physique, ni à aucune version naïve de la biologie évolutionniste. Comme nous venons de le voir, elle est compatible avec les mécanismes détaillés par les sciences physiques et biologiques, mais elle fait l'objet d'une *autre* science naturelle spécialisée. A mon avis, se méfier d'un pareil projet d'une sémantique naturelle à cause de l'impossibilité de sa réduction aux lois de la physique serait comme se méfier des règles de priorité dans l'art de naviguer à cause de l'impossibilité de trouver une traduction moléculaire de termes comme « babord amûres ».

CONCLUSION

Ainsi les atomes sémantiques possèdent une structure interne spécifique, avec des composantes cruciales, invariables, et des composantes secondaires, variables d'un locuteur à l'autre. Ne pas saisir cette différence entre les composantes, ou confondre les unes avec les autres, signifie ne pas « posséder » une certaine signification, ne pas posséder un certain concept, ne pas être connecté avec le monde de façon sémantiquement correcte en ce qui concerne *ce* concept, *ce* terme lexical. Notre capacité à saisir ces composantes constitue la base naturelle de la causalité sémantique. Ce processus de sélection sémantique, il faut le penser répété de façon spécifique pour chacun de nos concepts fondamentaux, c'est-à-dire, pour chacun des termes de notre lexique. Il s'agit, au plus, d'un répertoire de quelques centaines de milliers de termes. Un répertoire qui est ouvert à de nouvelles insertions, pourvu que de fortes contraintes structurales soient respectées. Chaque terme possède sa spécificité et les différentes langues ont des moyens un peu différents de

sélectionner les composantes cruciales et les composantes secondaires, en créant des répertoires lexicaux légèrement différents. Ces différences sont *toujours* limitées par des contraintes universelles, ce qui veut dire que seul un sous-ensemble de tout ce qui nous est concevable est aussi monolexicalisable.

Jusqu'à ces dernières années, les différences lexicales entre les langues (les mots « intraduisibles ») ont retenu l'attention des linguistes beaucoup plus que les profondes et remarquables similitudes entre tous les lexiques de toutes les langues naturelles. Cette similitude, étudiée à un niveau de profondeur morpho-syntaxique et sémantique adéquat, offre déjà quelques généralisations scientifiques intéressantes.

Ces généralisations ne sont que partielles et encore provisoires, mais elles démontrent déjà que nous sommes *naturellement* équipés pour saisir ces nombreuses structures spécifiques, sur la base de données limitées, épisodiques et apparemment largement insuffisantes. L'ensemble de nos systèmes langagiers et beaucoup de nos théories spontanées du monde sont mobilisés par le processus d'acquisition. Ces différents systèmes linguistiques et cognitifs, en grande partie modulaires, restent distincts dans leurs principes fonctionnels et architecturaux, mais ils interagissent étroitement et finement dans l'acquisition du lexique.

Qu'il y ait autant de structures spécifiques *fines* qu'il y a de significations monolexicales (ou concepts « primitifs ») *possibles* ne devrait pas nous étonner. Un ordre de grandeur de quelques centaines de milliers, ou quelques millions, de structures spécifiques pour un répertoire biologique est parfaitement normal (Piattelli-Palmarini, 1986, 1989). Je me permets de souligner que chacun de nos organismes possède à la naissance un bien plus grand nombre de *types* d'anticorps. Il s'agit de structures de haute complexité, qui sont tout aussi spécifiques, tout aussi diversifiées, et tout aussi innées, que nos concepts dits primitifs.

Massimo PIATTELLI-PALMARINI**,
Center for Cognitive Science, M.I.T.,
Cambridge, Massachusetts.

** Je remercie François Dell, Chantal Hunt et Jean-Michel Roy pour leur lecture critique d'une première version de ce manuscrit.

BIBLIOGRAPHIE

BAKER (Mark), 1988, *Incorporation : A Theory of Grammatical Function Changing,* Chicago, Ill., Chicago University Press.

BOLTZMANN (Ludwig), 1904, « On a Thesis of Schopenhauer », *in* B. McGUINNESS, ed., 1974, *Theoretical Physics and Philosophical Problems,* Dordrecht, Holland, D. Reidel.

CAMPBELL (Donald T.), 1974, « Evolutionary Epistemology », *in* P.A. SCHILPP, ed., *The Philosophy of Karl Popper,* La Salle, Ill., Open Court, p. 413-463.

CAREY (Susan), 1985, *Conceptual Change in Childhood,* Cambridge, MA, Bradford Books/MIT Press.

CARTER (Richard J.), 1980, « La notion d'explication en sémantique », *Langue française, 46.*

CARTER (Richard J.), 1984a, « Compositionality and Polysemy », *in* LEVIN & TENNY, eds, 1988, p. 167-204.

CARTER (Richard J.), 1984b, « On Movement », *in* LEVIN & TENNY, eds, 1988, p. 231-252.

CARTER (Richard J.), 1984c, « Sous-catégorisation et régularités sélectionnelles », *Communications, 40,* p. 181-209.

CHOMSKY (Noam), 1955, « Transformational Analysis », University of Pennsylvania, Thèse de doctorat, *in* CHOMSKY, 1985.

CHOMSKY (Noam), 1959, « A Review of B.F. Skinner's 'Verbal Behavior' », *Language, 35/*1, p. 26-58.

CHOMSKY (Noam), 1981, *Lectures on Government and Binding : The Pisa Lectures,* Dordrecht, Holland, Foris.

CHOMSKY (Noam), 1985, *The Logical Structure of Linguistic Theory,* Chicago, Ill., The University of Chicago Press.

CHOMSKY (Noam), 1986, *Knowledge of Language : Its Nature, Origin, and Use,* Convergence, New York, Praeger Scientific.

DENNETT (Daniel C.), 1983, « Intentional Systems in Cognitive Ethology : The 'Panglossian Paradigm' Defended », *The Behavioral and Brain Sciences, 6/*3, p. 343-390.

DENNETT (Daniel C.), 1987, *The Intentional Stance,* Cambridge, MA, Bradford Books/MIT Press.

DENNETT (Daniel C.), 1988, « Ways of Establishing Harmony », *in* B. MCLAUGHLIN, ed., *Essays in the Honour of Fred Dretske* (en préparation).

DOWTY (David R.), 1979, *Word Meaning and Montague Grammar,* Dordrecht, Holland, D. Reidel.

DRETSKE (Fred), 1981, *Knowledge and the Flow of Information,* Cambridge, MA, Bradford Books/MIT Press.

DRETSKE (Fred), 1986, « Misrepresentation », *in* R. BODGAN, ed., *Belief,* Oxford, Oxford University Press.

DRETSKE (Fred), 1988, *Explaining Behaviour : Reasons in a World of Causes,* Cambridge, MA, Bradford Books/MIT Press.

ELMAN (Jeffrey L.), 1989, « Representation and Structure in Connectionist Models », in *Center for Research in Language Technical Report 8903,* La Jolla, CA, University of California, San Diego.

FILLMORE (Charles), 1968, « The Case for Case », *in* E. BACH & R. HARMS, eds, *Universals in Linguistic Theory,* New York, Holt, Rinehart & Winston, p. 1-88.

FODOR (Jerry A.), 1970, « Three Reasons for not Deriving 'Kill' from 'Cause to Die' », *Linguistic Inquiry, 1,* p. 429-438.

FODOR (Jerry A.), 1987, *Psychosemantics,* Cambridge, MA, Bradford Books/MIT Press.

FODOR (Jerry A.), 1989, « A Theory of Content », Department of Philosophy, City University of New York, manuscrit non publié.

FODOR (Jerry A.) & PYLYSHYN (Zenon), 1988, « Connectionism and Cognitive Architecture : A Critical Analysis », *in* S. PINKER & J. MEHLER, eds, *Connections and Symbols,* Cambridge, MA, Bradford Books/MIT Press.

GLEITMAN (Lila R.), 1986, « Biological Dispositions to Learn Language », *in* W. DEMOPOULOS & A. MARRAS, eds, *Language Learning and Concept Acquisition : Foundational Issues,* Norwood, NJ, Ablex.

GOLDSTONE (Robert L.), GENTNER (Derdre) & MEDIN (Douglas L.), 1989, « Relations Relating Relations », in *11th. Annual Conference of the Cognitive Science Society,* Ann Arbor, Michigan, Lawrence Erlbaum Associates, vol. XI, p. 131-138.

GRIMSHAW (Jane), 1979, « Complement Selection and the Lexicon », *Linguistic Inquiry, 10/2.*

GRIMSHAW (Jane), 1982, « On the Lexical Representation of Romance Reflexive Clitics », *in* J. BRESNAN, ed., *The Mental Representation of Grammatical Relations,* Cambridge, MA, MIT Press.

HALE (Kenneth) & KEYSER (Samuel Jay), 1988, « Explaining and Constraining the English Middle », *in* C. TENNY, ed., *Studies in Generative Approaches to Aspect,* Lexicon Project Working Papers, n° 24, Cambridge, MA, MIT Center for Cognitive Science, p. 41-57.

HALE (Kenneth) & KEYSER (Samuel Jay), 1989, « The Syntactic Character of Thematic Structure », manuscrit non publié, Lexicon Project, MIT Center for Cognitive Science.

HIGGINBOTHAM (James T.), 1985, « On Semantics », *Linguistic Inquiry, 16/4,* p. 547-593.

HIGGINBOTHAM (James T.), 1986, « Elucidations of Meaning », *Linguistics and Philosophy, 12/3.*

HIGGINBOTHAM (James T.), 1988, « Knowledge of Reference », *in* A. GEORGE, ed., *Reflections on Chomsky,* Oxford, Basil Blackwell.

JACKENDOFF (Ray), 1983, *Semantics and Cognition,* Cambridge, MA, MIT Press.

JACKENDOFF (Ray), 1987, « The Status of Thematic Relations in Linguistic Theory », *Linguistic Inquiry, 18/3.*

KEIL (Frank C.), 1979, *Semantic and Conceptual Development : An Ontological Perspective,* Cambridge, MA, Harvard University Press.

KAYNE (Richard S.), 1975, *French Syntax,* Cambridge, MA, MIT Press.

KRIPKE (Saul), 1972, *Meaning and Necessity,* Oxford, Oxford University Press.

LANDAU (Barbara) & GLEITMAN (Lila R.), 1985, *Language and Experience : Evidence from the Blind Child,* Cambridge, MA, Harvard University Press.

LASNIK (Howard) & URIAGEREKA (Juan), 1988, *A Course in GB Syntax : Lectures on Binding and Empty Categories,* Cambridge, MA, MIT Press.

LEDERER (Anne), GLEITMAN (Henry) & GLEITMAN (Lila R.), 1989, « Syntactic Bootstrapping : Are the Data for a Deductive, Principle-Driven Verb Learning Procedure Available to Children ? », communication présentée à *The 14th. Annual Conference on Language Development,* Boston University, oct. 1989.

LEVIN (Beth) & TENNY (Carol), eds, 1988, *On Linking : Papers by Richard Carter,* Lexicon Project Working Papers, n° 25, Cambridge, MA, MIT Center for Cognitive Science.

LOAR (Brian), 1981, *Mind and Meaning,* Cambridge, U.K., Cambridge University Press.

LORENZ (Konrad), 1941, « Kants Lehre vom Apriorischen im Lichte gegenwärtiger Biologie », *Blätter für deutsche Philosophie, 15,* p. 94-125.

LORENZ (Konrad), 1982, « Kants Doctrine of the A Priori in the Light of Contemporary Biology », trad. de LORENZ, 1941, *in* H.C. PLOTKIN, ed., *Learning, Development and Culture : Essays in Evolutionary Epistemology,* New York, John Wiley & Sons.

MACNAMARA (John), 1982, *Names for Things : A Study of Human Learning,* Cambridge, MA, Bradford Books/MIT Press.

MARKMAN (Ellen M.), 1989, *Categorization and Naming in Children : Problems of Induction,* Cambridge, MA, Bradford Books/MIT Press.

MARSLEN-WILSON (William) & TYLER (Lorraine), 1980, « The Temporal Structure of Spoken Language Understanding », *Cognition, 8,* p. 1-71.

MAY (Robert), 1985, *Logical Form : Its Structure and Derivation,* Cambridge, MA, MIT Press.

MCCLOSKEY (Michael), 1983, « Naive Theories of Motion », *in* D. GENTNER & A.L. STEVENS, eds, *Mental Models,* Hillsdale, NJ, Lawrence Erlbaum.

MILLER (George A.), 1986, « Dictionaries in the Mind », *Language and Cognitive Processes, 1*/3, p. 171-185.

MILLER (George A.) & GILDEA (Patricia M.), 1987, « How Children Learn Words », *Scientific American, 257*/3, p. 94-99.

MILLIKAN (Ruth Garrett), 1984, *Langage, Thought, and Other Biological Categories,* Cambridge, MA, Bradford Books/MIT Press.

MILLIKAN (Ruth Garrett), 1986, « Thoughts without Laws : Cognitive Science with Content », *Philosophical Review, 95*/1, p. 47-80.

MILLIKAN (Ruth Garrett), 1989, « Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox », *Philosophical Review,* à paraître en 1990.

NERSESSIAN (Nancy J.) & RESNICK (Lauren B.), 1989, « Comparing Historical and Intuitive Explanations of Motion : Does 'Naive Physics' Have a Struc-

ture ? », in *11th. Annual Conference of the Cognitive Science Society,* Ann Arbor, Michigan, Lawrence Erlbaum Associates, p. 412-418.

PEIRCE (Charles Sanders), 1896, « The Scientific Attitude and Fallibilism », *in* J. BUCHLER, ed., 1955, *Philosophical Writings of Peirce,* New York, Dover.

PESETSKY (David), 1985, « Morphology and Logical Form », *Linguistic Inquiry, 16/2,* p. 193-246.

PESETSKY (David), 1987, « Binding Problems with Experiencer Verbs », *Linguistic Inquiry, 18,* p. 126-140.

PETITOT (Jean), 1985, *Morphogenèse du sens.* I, Paris, Presses universitaires de France.

PETITOT (Jean), 1989, « Hypothèse localiste, modèles morphodynamiques et théories cognitives : remarques sur une note de 1975 », à paraître dans *Semiotica.*

PIATTELLI-PALMARINI (Massimo), 1985, « The Waning of Parsimony », *Scientia,* numéro spécial « La vita e la sua storia », sous la dir. de L. BULLINI, M. FERRAGUTI, A. OLIVERIO & F. MONDELLA, p. 265-279.

PIATTELLI-PALMARINI (Massimo), 1986, « The Rise of Selective Theories : A Case Study and some Lessons from Immunology », *in* W. DEMOPOULOS & A. MARRAS, eds, *Language Learning and Concept Acquisition : Foundational Issues,* Norwood, NJ, Ablex.

PIATTELLI-PALMARINI (Massimo), 1988 a, « Can We Rationally Believe Something We Do Not Quite Understand ? », *in* Actes du colloque *Gli Stili dell' Argomentazione,* Modena, Fondazione San Carlo, à paraître.

PIATTELLI-PALMARINI (Massimo), 1988 b, « Not on Darwin's Shoulders : A Critique of Evolutionary Epistemology », in *Boston Colloquia for the Philosophy of Science,* Boston, MA, à paraître.

PIATTELLI-PALMARINI (Massimo), 1989, « Evolution, Selection and Cognition : From 'Learning' to Parameter Setting in Biology and in the Study of Language », *Cognition, 31/1,* p. 1-44.

PINKER (Steven), 1989, *Learnability and Cognition : the Acquisition of Argument Structure,* Cambridge, MA, MIT Press.

POPPER (Karl R.), 1972, *Objective Knowledge : An Evolutionary Approach,* Oxford, Clarendon Press.

PUSTEJOVSKY (James), 1988, « The Geometry of Events », *in* C. TENNY, ed., *Studies in Generative Approaches to Aspect,* Cambridge, MA, MIT Center for Cognitive Science, p. 19-39.

PUTNAM (Hilary), 1970, « Is Semantics Possible ? », *in* H. KIEFER & M. MUNITZ, eds, *Languages, Belief and Metaphysics,* New York, State University of New York.

PUTNAM (Hilary), 1975, « The Meaning of 'Meaning' », *in* K. GUNDERSON, ed., *Language, Mind and Knowledge,* Minneapolis, Minnesota, University of Minnesota Press.

QUINE (Willard Van Orman), 1960, *Word and Object,* Cambridge, MA, MIT Press.

QUINE (Willard Van Orman), 1974, *The Roots of Reference (The Paul Carus Lectures),* La Salle, Ill., Open Court.

RAPPAPORT (Malka) & LEVIN (Beth), 1988, « What to Do with Theta-Roles », *in* W. WILKINS, ed., *Thematic Relations,* New York, Academic Press.

RIEMSDIJK (Henk Van) & WILLIAMS (Edwin), 1986, *Introduction to the Theory of Grammar,* Cambridge, MA, MIT Press.

ROEPER (Thomas) & WILLIAMS (Edwin), eds, 1987, *Parameter Setting,* Dordrecht, Netherlands, D. Reidel.

SALASOO (A.) & PISONI (David B.), 1985, « Interaction of Knowledge Sources in Spoken Word Identification », *Journal of Memory and Language, 24,* p. 210-231.

SEGAL (Gabriel) & SOBER (Elliott), 1989, « The Causal Efficacy of Content », manuscrit non publié, Madison, University of Wisconsin, Department of Philosophy.

SMITH (Edward E.) & MEDIN (Douglas L.), 1981, *Categories and Concepts,* Cambridge, MA, Harvard University Press.

SPELKE (Elisabeth S.), 1985, « Perception of Unity, Persistence and Identity : 'Thoughts on Infants' Conception of Objects », *in* J. MEHLER & R. FOX, eds, *Neonate Cognition : Beyond the Blooming Buzzing Confusion,* Hillsdale, NJ, Lawrence Erlbaum.

SPELKE (Elisabeth S.), 1988, « Where Perceiving Ends and Thinking Begins : The Apprehension of Objects in Infancy », *in* A. YONAS, ed., *Perceptual Development in Infancy,* Minneapolis, MI, University of Minnesota Press.

TALMY (Leonard), 1985, « Lexicalization Patterns », *in* T. SHOPEN, ed., *Language Typology and Syntactic Description,* Cambridge, U.K., Cambridge University Press.

TANENHAUS (Michael K.), GARNSEY (Susan M.) & BOLAND (Julie), 1989, « Combinatory Lexical Information and Language Comprehension », *in* G. ALTMAN, ed., *Cognitive Models of Speech Processing : Psycholinguistic and Computational Perspectives,* Cambridge, MA, MIT Press.

TENNY (Carol), ed., 1988a, *Studies in Generative Approaches to Aspect,* Lexicon Project Working Papers, n° 24, Cambridge, MA, MIT Center for Cognitive Science.

TENNY (Carol), 1988b, « The Aspectual Interface Hypothesis », *in* C. TENNY, ed., 1988a, p. 1-18.

CONNEXIONNISME ET COGNITION :
À LA RECHERCHE DES BONNES QUESTIONS

Lorsque les réseaux de neurones formels apportèrent la bonne nouvelle connexionniste aux quatre coins du petit monde des sciences cognitives[1], ils se présentaient sous des dehors fort différents des systèmes classiques, ceux que l'intelligence artificielle construisait et à l'image desquels les sciences cognitives tentaient de façonner leurs modèles[2]. Si

1. Le connexionnisme d'aujourd'hui est la résurgence d'un courant qui a précédé, et dont est issu le mouvement « cognitiviste » majoritaire depuis une trentaine d'années dans les sciences cognitives et l'Intelligence Artificielle. Le premier connexionnisme est issu des travaux de Warren McCulloch et Walter Pitts et constitue l'une des plus durables contributions de la cybernétique (cf. « A Logical Calculus of the Ideas Immanent in Nervous Activity », leur article de 1943 repris *in* Warren S. McCULLOCH, *Embodiments of Mind,* Cambridge, MA, MIT Press, 1965/1988). L'un de ses prolongements les mieux connus est l'invention par F. Rosenblatt du perceptron (cf. Frank ROSENBLATT, *Principles of Neurodynamics,* New York, Spartan, 1962). On s'accorde pour voir dans le livre de Marvin MINSKY et Seymour PAPERT, *Perceptrons,* Cambridge, MA, MIT Press, 1969 (nouv. éd. 1989), le certificat de décès sociologique du premier connexionnisme. Des éléments d'histoire de ce mouvement sont fournis dans les *Cahiers du CREA, 6* et *7,* nov. 1985 (épuisés mais consultables au CREA, 1, rue Descartes, 75005 Paris). L'essor du cognitivisme est ponctué de nombreux écrits ; on en trouvera une description partielle ci-dessous ; des exposés plus détaillés mais accessibles se trouvent notamment dans John HAUGELAND, ed., *Mind Design,* Cambridge, MA, MIT Press, 1981 ; Daniel ANDLER, « Les sciences de la cognition », in *La Philosophie des sciences aujourd'hui,* sous la dir. de Jean HAMBURGER, Paris, Gauthier-Villars, 1986 ; ID., « Progrès en situation d'incertitude », *Le Débat, 47,* nov-déc. 1987, p. 213-234 ; ID., article « Cognitives (sciences) », *Encyclopaedia Universalis,* Paris, nouv. éd., 1989. Le (néo) connexionnisme s'est développé à partir de la fin des années 1970, et a acquis une notoriété considérable grâce à la publication en 1986 d'un ouvrage en deux forts volumes, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition,* par David RUMELHART, James McCLELLAND et le PDP Research Group, Cambridge, MA, MIT Press. (Un chapitre en a partiellement été traduit dans le n° 47 du *Débat.*) L'anthologie de James ANDERSON et Edward ROSENFELD, *Neurocomputing - Foundations of Research,* Cambridge, MA, MIT Press, 1988, permet de suivre très précisément le développement du connexionnisme, des origines à nos jours. Une très éclairante analyse des fondements de l'approche PDP se trouve *in* Andy CLARK, *Microcognition. Philosophy, Cognitive Science, and Parallel Distributed Processing,* Cambridge, MA, MIT Press, 1989.

2. On peut prendre très au sérieux une conception de l'esprit inspirée du modèle de l'ordinateur sans pour cela estimer que l'Intelligence Artificielle contribue notablement au progrès des sciences cognitives. Telle est, par exemple, la position de Jerry FODOR, in *The*

différents, de fait, qu'il fallait se frotter les yeux plusieurs fois avant de pouvoir admettre que les connexionnistes et leurs concurrents cognitivistes (nous les appellerons aussi, suivant l'exemple de leurs champions J. Fodor et Z. Pylyshyn[3], « les classiques ») se proposaient d'expliquer *les mêmes phénomènes.* Les débats qui s'ensuivirent modifièrent à ce point la perspective qu'on put penser que non seulement les uns et les autres parlaient de la même chose, mais qu'ils en proposaient finalement *le même genre d'explication* — que les réseaux n'étaient que des systèmes classiques, d'un genre sans doute un peu particulier, décrits d'une manière inhabituelle. Et il fallut se frotter les yeux à nouveau pour repérer malgré tout des différences. Aujourd'hui Fodor et Pylyshyn, défenseurs et illustrateurs du classicisme, prétendent placer le connexionnisme devant l'alternative suivante : ou bien les réseaux sont foncièrement différents des systèmes classiques, mais ils sont et demeureront alors inaptes à modéliser des aspects essentiels de la cognition ; ou bien ils peuvent surmonter leurs faiblesses actuelles, mais ils ne sont en fait alors que des systèmes classiques vus par le petit bout de la lorgnette.

Le débat oscille depuis le début entre deux extrêmes : les connexionnistes radicaux prétendent que leur *sujet* d'étude est le même que celui des classiques, mais que ceux-ci ont une conception erronée de la *nature* des phénomènes, conception qui les amène à un choix malheureux du niveau de description, donc à une caractérisation (et le cas échéant une simulation) inadéquate de la cognition ; les classiques intransigeants maintiennent que les connexionnistes se trompent de sujet (à la manière de géologues, par exemple, qui ne voudraient parler que de quanta), ce qui prive naturellement leurs explications et leurs simulations de toute véritable pertinence. Bref, nous serions sommés de trancher entre deux hypothèses : celle d'un sujet unique et de deux explications (dont une bonne et l'autre mauvaise) ; celle de deux sujets (dont l'un correspond au domaine pré-théoriquement visé et l'autre pas), chacun muni de son explication propre (et dont l'une seulement nous intéresse).

Peut-être est-il possible d'échapper à cette alternative. Le débat n'a, en tout cas, visiblement pas pour issue prochaine la désignation d'un vainqueur. Il permettra sans doute en revanche de faire la part de ce qui tient, entre les deux « camps »[4], du *désaccord,* et ce qui tient du *malentendu.*

Modularity of Mind, Cambridge, MA, MIT Press, 1983 ; tr. fr. par Abel GERSCHENFELD, *La Modularité de l'esprit,* Paris, Minuit, 1986.

3. Jerry FODOR, Zenon PYLYSHYN, « Connectionism and Cognitive Architecture : A Critical Analysis », *Cognition, 28,* 1988, p. 3-71. Repris *in* Steven PINKER, Jacques MEHLER, eds, *Connections and Mind,* Cambridge, MA, MIT Press, 1988.

4. Nombreux sont ceux qui rejettent l'idée de deux camps hostiles, pour diverses raisons qui ne tarderont pas à apparaître.

L'enjeu pour l'épistémologue est de discerner comment s'associent une *conception* de la nature véritable du domaine pré-théoriquement visé sous l'appellation « cognition », et un *type d'explication* des phénomènes qui le constituent.

I. — MACHINES

Le cognitivisme et le connexionnisme se définissent en partie (mais non, on le verra, exclusivement) par la référence à un type de machine. Pour les défenseurs du premier, il s'agit de l'ordinateur dit de von Neumann, pour ceux du second, du réseau de neurones formels (dit aussi réseau neuromimétique — en anglais, *neural net*). Nous présupposons la familiarité avec l'ordinateur, en tant qu'objet théorique, nous décrivons rapidement à présent le réseau[5].

Bien qu'il existe de nombreuses variétés de réseau, présentant de très importantes différences (alors que la machine de von Neumann est essentiellement unique), on peut, pour les besoins de la discussion, dresser une sorte de « portrait-robot » du réseau connexionniste.

Il s'agit d'un ensemble d'*automates* très simples interconnectés. Les *connexions* permettent à un automate tel que i de transmettre à un automate j une stimulation déterminée par l'état d'activité u_i de i et modulée par un *poids synaptique* w_{ji} ne dépendant que du canal. Le poids est affecté d'un signe : s'il est positif, la stimulation est positive (excitatrice) ; s'il est négatif, elle est négative (inhibitrice). Les automates (ou *unités*) sont en général tous identiques — ce sont souvent des automates *à seuil,* dont l'activité est soit 0 soit 1, et qui sont capables seulement de comparer la somme pondérée des stimulations afférentes $\Sigma_i u_i w_{ji}$ à un seuil s_j et de se mettre ou se maintenir en état d'activité ($u_i = 1$) si ce seuil est dépassé, de s'éteindre ou rester inactif ($u_i = 0$)

5. Toutes les références de la note 1 postérieures à 1985 contiennent des présentations plus ou moins détaillées des réseaux de neurones formels (nous dirons simplement désormais « réseaux »). En français, on dispose d'un manuel de Gérard WEISBUCH, *Dynamique des systèmes complexes. Une introduction aux réseaux d'automates,* Paris, Interéditions/Ed. du C.N.R.S., 1989 ; on pourra aussi consulter le numéro spécial de *Intellectica,* la revue de l'Association pour la recherche cognitive, de février 1990, dirigé par Daniel Memmi et Yves-Marie Visetti. En anglais, citons l'ouvrage collectif dirigé par Geoffrey E. HINTON et James A. ANDERSON, *Parallel Models of Associative Memory,* Hillsdale, NJ, Erlbaum, 1981, qui marque la renaissance du connexionnisme, et le magistral traité de Daniel AMIT, *Modeling Brain Function. The World of Attractor Neural Networks,* Cambridge, Cambridge University Press, 1989.

sinon. Le système est donc caractérisé, à chaque étape de son évolution, qui est discrète, par un *vecteur d'activation* u = (u_1, ..., u_n) ; la transition d'une étape à la suivante résulte d'une mise à jour, soit par tous les automates simultanément, soit par un seul choisi par exemple au hasard, de leur activité.

Lorsque le réseau est utilisé pour transformer une famille d'entrées en certaines sorties spécifiées ou spécifiables (en d'autres termes, lorsque l'on choisit de le faire fonctionner comme une fonction incarnée), le processus commence par l'imposition d'un certain vecteur d'activation u_0, qui peut être considéré comme la donnée ou *input*, se poursuit par itération de la règle de transition, et se termine (dans les cas favorables) lorsque le système atteint un équilibre, caractérisé par un vecteur $u_N = u_\infty$, résultat ou *output* de ce qu'on peut appeler le calcul effectué par le réseau. Ce n'est pas là le seul usage possible d'un réseau, ni peut-être même le plus intéressant, mais la discussion n'en exige pas davantage pour le moment[6].

En tant que calculateur abstrait (ou, si l'on veut, de système de traitement de l'information — mais il n'a pas encore été question d'information), le réseau présente, en première analyse, d'importantes différences avec la machine de von Neumann :

1. Dans celle-ci, le processus est « séquentiel », en ce sens que les opérations élémentaires sont effectuées l'une après l'autre ; dans un réseau, un grand nombre d'entre elles sont faites simultanément et indépendamment les unes des autres.

2. Un réseau est foncièrement homogène (même s'il n'est pas totalement connecté, c'est-à-dire que chaque unité n'influe que sur certaines autres), en ce sens qu'on n'y distingue pas, comme dans l'ordinateur classique, une hiérarchie de sous-systèmes spécialisés dans des tâches interdépendantes de complexité croissante.

3. En particulier, le processus n'est dirigé dans le réseau par rien qui ressemble à une unité centrale de contrôle — c'est tout le contraire de l'ordinateur.

4. Ce qui fait la spécificité d'un réseau, outre le nombre de ses unités, c'est la matrice des poids synaptiques. C'est de ce vecteur que dépend, à *input* égal, le comportement du réseau et le résultat de son calcul — en ce sens il constitue sa « compétence ». C'est donc l'équivalent du programme de l'ordinateur, mais c'est tout différent — aussi différent, par exemple, que l'ensemble des forces agissant sur une bulle de savon peut l'être d'une suite d'instructions pour résoudre le système d'équations différentielles déterminant la position d'équilibre de la bulle.

---

6. Les termes de cette présentation élémentaire sont largement empruntés à mon article de l'*Encyclopaedia Universalis*, *art. cit. supra* n. 1.

5. Enfin, un réseau peut, en théorie, traiter des grandeurs continues, alors que l'ordinateur se nourrit d'entités discrètes (il est *« digital »*, dit l'anglais, ce qu'on nous oblige à traduire par « numérique » — il faudrait dire « chiffral », ce qui éviterait bien des confusions...).

Nous nous interrogerons bientôt sur la solidité de ces contrastes, mais il nous faut d'abord comprendre comment, dans chacun des deux « paradigmes » en présence, la machine devient un système cognitif en puissance.

II. — SYSTÈMES COGNITIFS

a. Systèmes classiques

Le cognitivisme classique a pour postulat fondamental le *fonctionnalisme.* Ce terme recouvre en fait un écheveau de doctrines aux ramifications souvent subtiles, élaborées dans le cadre général du problème corps-esprit[7]. Il est nécessaire ici de n'en retenir que le principe d'une *double description* des systèmes cognitifs. Ils doivent être vus à la fois comme des systèmes matériels et comme des systèmes informationnels[8], et seule importe la possibilité de principe d'un passage d'une description à l'autre. Cette possibilité une fois établie, la tâche des sciences cognitives se borne à caractériser les systèmes informationnels capables d'exhiber un comportement conforme aux principaux aspects observables ou inférables de notre vie mentale. Tout le reste est question d'intendance, c'est-à-dire d' « implémentation ». Symétriquement, la stratégie de l'Intelligence Artificielle classique est de modeler la machine de von Neumann, pour laquelle le passage entre les deux descriptions est assez bien balisé, en sorte d'en tirer des effets d'intelligence (ou plus généralement peut-être, mais plus obscurément aussi, des effets cognitifs ou mentaux).

Un système informationnel classique est composé de deux parties. La partie *variable* est un ensemble de représentations. La partie *fixe* (invariable) est un ensemble de dispositifs capables d'effectuer sur les représentations certaines transformations. En première approximation, les représentations constituent les « connaissances » du système, et les transformations font évoluer ces connaissances, en les combinant entre elles

7. Une excellente introduction à la question est fournie par Pierre JACOB, « Le problème du rapport du corps et de l'esprit aujourd'hui. Essai sur les forces et les faiblesses du fonctionnalisme », à paraître dans *Approches de la cognition,* sous la dir. de D. ANDLER, Paris, Gallimard, « Folio ».

8. J'utilise ce terme ici dans un sens aussi neutre que possible, sans vouloir le distinguer par exemple de « système cognitif ».

d'une part, en les modifiant en fonction d'informations nouvelles d'autre part.

Les représentations sont des expressions d'un langage interne du système, langage du genre de ceux de la logique formelle[9]. Elles désignent notamment des entités, des situations, des événements singuliers, ainsi que des relations générales entre entités, situations, événements. Ce qui est décrit appartient au monde extérieur — à l'environnement au sens le plus large —, mais aussi le cas échéant au monde intérieur du système, qui peut avoir de lui-même une certaine représentation.

Les transformations sont pour l'essentiel des inférences : elles partent typiquement d'un ensemble de représentations de la forme (A, A → B) et en font (B), ou encore (A, A → B, B). Ce sont naturellement des opérations formelles, qui peuvent aussi bien être vues comme des fonctions récursives de nombres entiers codant les formules — ce codage se compose avec la représentation —, en sorte que tel fait est maintenant représenté par un nombre entier plutôt que par une formule. A cette légère modification de point de vue ne correspond aucun changement matériel : en dernière analyse, une machine de Turing ne fait qu'inscrire et effacer des croix dans les cases d'un ruban, croix qui ne sont intrinsèquement pas davantage des nombres que des formules ou que des propositions ! Mais on comprend facilement alors qu'une machine de Turing universelle[10], capable par définition d'effectuer toute opération réalisable par n'importe quelle autre machine de Turing, puisse effectuer toute opération « qui peut être décrite de manière exhaustive et dépourvue d'ambiguïté, tout ce que des mots peuvent exprimer complètement et sans ambiguïté »[11], qu'il s'agisse de nombres ou de formules. On comprend aussi du même coup que toute opération cognitive puisse être vue comme une suite d'étapes élémentaires dont chacune est une inférence au sens faible d'étape d'un calcul correct[12].

Pour essentielle et apparemment simple que soit sur le plan abstrait la

9. On parle souvent dans ce contexte de langages *symboliques*, comme s'il en était qui ne le sont pas ! Le terme « symbole » renvoie ici d'une part à la logique dite symbolique, d'autre part à la matérialisation au moins potentielle dans un système de transformations de systèmes de symboles. On est donc assez loin de l'usage courant.

10. Et donc, moyennant les idéalisations habituelles (temps et mémoire illimités), une machine de von Neumann.

11. Selon les termes de von Neumann au Hixon Symposium, le 29 septembre 1948, rapportés par Hermann GOLDSTINE, *The Computer from Pascal to von Neumann*, Princeton, Princeton University Press, 1972, p. 276. Von Neumann parlait, en fait, des réseaux de McCulloch et Pitts.

12. Il n'est donc pas question de réduire la cognition à la logique déductive classique *au niveau des représentations*. Ce point délicat est la source de malentendus sans fin et de critiques sans fondement.

distinction entre partie fixe ou processuelle et partie variable ou factuelle, elle soulève une difficulté. Pour l'apercevoir, il suffit de substituer aux termes inhabituels qui viennent d'être utilisés ceux, plus courants, de « programme » et de « données ». En effet, si l'on se réfère à la spécification physique d'un ordinateur, seul son plan de câblage est fixe, et correspond à la partie fixe du système informationnel qu'il constitue ou sous-tend. Tout le reste est variable : on sait que le trait de génie des inventeurs de l'ordinateur moderne fut de donner aux règles de fonctionnement le même statut qu'aux entités sur lesquelles elles opèrent : la différence entre programme et données, en ce qui concerne les systèmes matériels qui correspondent — ou devraient correspondre — aux systèmes cognitifs que l'on considère, n'existe que dans le regard de l'observateur ; les machines matérielles qui correspondent aux systèmes cognitifs *intéressants* sont déjà elles-mêmes virtuelles. Entre le niveau de la machine « réellement » matérielle (l'ordinateur non encore programmé) et celui de la machine cognitive s'insèrent donc des niveaux intermédiaires au statut moins clair sur le plan cognitif que sur le plan informatique[13]. Cette difficulté a pour effet, on le verra, de compliquer la comparaison entre systèmes classiques et connexionnistes.

Mais venons-en à la question centrale du rapport entre les niveaux matériel et informationnel ou cognitif. Ce rapport résulte d'une *double* articulation. La première est constituée par le rapport entre la syntaxe et la sémantique du langage formel dans lequel s'expriment les représentations internes du système. Les énoncés que sont ces représentations ont, on le sait, deux visages : leur structure morphologique détermine les transformations syntaxiques auxquelles ils sont sujets en vertu des règles d'inférence du langage, tandis que leur interprétation dans un univers donné (par exemple, l'environnement du système) leur donne une valeur sémantique qu'on peut assimiler en première analyse à une proposition portant sur les objets qui sont l'interprétation des termes du langage. Le parallélisme entre syntaxe et sémantique, que garantit le théorème de complétude de Gödel, explique que lorsque le système passe, en vertu de la syntaxe, d'une représentation R à une représentation R', il passe en même temps d'une proposition vraie (ou supposée telle par le système) à une autre. Ainsi s'explique que le système reste « en contact » avec la réalité externe, tout en ne se guidant que sur ses représentations internes[14].

---

13. Entre les différents niveaux de description de l'ordinateur programmé le rapport est de « compilation ». Dans « Le paradigme de la compilation », in *Approches de la cognition, op. cit. supra* n. 7, Jean-Pierre DESCLÉS plaide précisément pour une réduction du rapport entre niveau matériel et niveau informationnel au rapport de compilation.

14. Il reste à comprendre comment il se fait que le système parte « du bon pied », c'est-à-dire avec des représentations atomiques fidèles. C'est là l'un des points les plus faibles du

Cette première articulation ne permet pas cependant de quitter le niveau informationnel ; elle n'assure pas à elle seule le passage au niveau matériel. Une seconde articulation est nécessaire, reliant cette fois la syntaxe et la physique, ou si l'on veut l'inférence et la cause. Cette articulation est réalisée par l' « incarnation » du calcul inférentiel dans un calculateur matériel — les détails étant sans pertinence. Le modèle réduit en est la calculatrice de poche qui « réalise » les lois de l'arithmétique ; le modèle en vraie grandeur, l'ordinateur, qui « réalise » les lois de la machine de Turing, ou encore un ensemble d'opérations permettant d'engendrer toute fonction récursive. Le principe est de modéliser les traits syntaxiques des formules par des propriétés physiques particulières de constituants d'une machine réelle, et de faire coïncider les lois de transition de la machine telles qu'elles s'expriment au niveau de ces propriétés particulières avec les règles de la syntaxe. C'est ainsi finalement qu'une machine concrète fonctionnant selon les lois de la physique peut être, d'une part, une « machine syntaxique » (peut se conformer aux règles d'une syntaxe formelle), et, d'autre part, selon l'expression de D. Denett, une « machine sémantique ».

Il ne reste plus qu'à souligner que lorsque les cognitivistes classiques caractérisent la cognition comme calcul sur des représentations, ils font référence à une notion parfaitement circonscrite de calcul, celle de Church et de Turing : une fonction calculable est — moyennant codage — identique à une fonction récursive sur les nombres entiers.

b. Systèmes connexionnistes

Il existe, on l'a dit, non pas un seul type de réseau connexionniste, mais une grande variété. Mais ce n'est pas là la seule ni la plus importante raison de distinguer plusieurs formes de connexionnisme. Des différences plus déterminantes encore se situent dans la façon dont ces réseaux sont vus comme systèmes cognitifs. Aussi sera-t-il difficile, parfois impossible, d'opposer à chaque aspect de la conception cognitiviste une position connexionniste unique et précise.

Pour fixer les idées du lecteur profane, voici deux exemples de systèmes connexionnistes. Le premier est un produit typique de l'approche dite PDP[15]. Dû à D. Rumelhart et J. McClelland[16], il est capable

---

cognitivisme classique, comme le reconnaît notamment J. Fodor ; cf. « Fodor's Guide to Mental Representation : The Intelligent Auntie's Vade-Mecum », *Mind*, vol. 44, 1985, p. 76-100, notamment les dernières lignes ; et sa tentative de solution dans *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, MIT Press, 1987.

15. Pour *Parallel Distributed Processing*, cf. *supra* n. 1.

16. *Op. cit. supra* n. 1, vol. 2, chap. 18, p. 216-271 : « On learning the past tenses of English verbs ».

d'apprendre à former le prétérit de tout verbe anglais à partir de l'infinitif. Les entrées comme les sorties sont des représentations phonémiques ; l'apprentissage est « supervisé » : un corpus d'exemples formés de couples infinitif/prétérit (*eat/ate* ; *be/was* ; *love/loved,* etc.) est d'abord montré au système — plus précisément, l'infinitif est présenté, puis la réaction spontanée du système est graduellement corrigée jusqu'à ce qu'elle soit conforme. Cette rectification progressive suit un algorithme relativement complexe, mais dont le principe général est celui formulé par Donald Hebb[17] : renforcer les poids synaptiques entre unités qui doivent être actives ou inactives simultanément, et réduire les poids dans la situation inverse. L'algorithme est indépendant de la fonction que le système doit apprendre, et son application n'exige pas l'intervention du modélisateur (c'est vraiment un algorithme[18] !). Quant au corpus, il est vaste, mais non exhaustif : une proportion non négligeable de verbes, tant réguliers qu'irréguliers, n'y figure pas. Le système est capable de maîtriser le corpus, au terme d'une (longue) période d'apprentissage ; après quoi il conjugue aussi presque infailliblement tout autre verbe anglais. Il est essentiel de remarquer qu'aucune règle n'est enseignée ou indirectement fournie au système par le modélisateur (en revanche, celui-ci est entièrement responsable du « pré-traitement » conduisant à la représentation phonémique, ainsi que du choix du corpus et du protocole d'apprentissage[19]).

Bon nombre de modèles de l'école PDP partagent avec celui-ci les caractéristiques suivantes. La tâche consiste à compléter une configuration dont l'environnement ne fournit qu'une partie. Dans un cas particulier fréquent, l'environnement fournit un vecteur x et le système doit compléter par f(x). Au contact d'un grand nombre d'exemples de configurations complètes (dans l'exemple, de points du graphe de la fonction f), le système s'adapte aux régularités de l'environnement en ajustant ses poids synaptiques, ce qui lui permet d'une part de réagir sans aucune erreur aux exemples qui lui ont été présentés au cours de l'apprentissage, d'autre part de réagir « intelligemment » à d'autres configurations incomplètes — soit en les assimilant à des parties de configurations connues, soit en y discernant un mélange de configurations connues, et en les complétant en conséquence. Bref, le système se comporte en détecteur de régularités statistiques multidimensionnelles.

---

17. Donald HEBB, *The Organization of Behavior,* New York, Wiley & Sons, 1949.

18. Ce qui laisse toutefois entier le problème de savoir quel genre d'ordinateur naturel l'exécuterait. Ce n'est pas (dans l'état actuel) le réseau lui-même.

19. Ce qui a valu à ce modèle de graves critiques, notamment de la part de Steven PINKER et Alan PRINCE, « On Language and Connectionism. Analysis of a Parallel Distributed Processing Model of Language Acquisition », *Cognition,* vol. 28, 1988, p. 73-193 ; et *in* S. PINKER, J. MEHLER, *op. cit. supra* n. 3.

Parmi les autres courants qui coexistent au sein du connexionnisme, celui qu'animent les physiciens joue un rôle privilégié ; l'un d'entre eux, D. Amit, a proposé de le désigner par le sigle ANN (pour *attractor neural network*)[20]. Notre deuxième exemple est le modèle de mémoire associative « adressable par le contenu »[21] proposé par John Hopfield dans un article mémorable[22] qui a donné au connexionnisme renaissant une impulsion décisive, et marque la naissance du courant ANN. Les réseaux de Hopfield, contrairement aux réseaux PDP, qui sont des perceptrons généralisés de type « feed forward » dans lesquels l'information se propage de manière unidirectionnelle, ménagent des connexions dans toutes les directions : ils sont complètement connectés (ou presque). Ils sont ainsi dotés d'une dynamique autonome déterminée par les poids synaptiques et caractérisée par des attracteurs. Chacun de ces attracteurs peut être considéré comme une « mémoire » (un souvenir) du réseau : toute stimulation suffisamment proche d'un attracteur place le système sur une orbite qui se stabilise en cet attracteur. Le problème est de savoir à quelles conditions le système peut, par un choix approprié de poids synaptiques, adopter pour attracteurs un certain ensemble prédéterminé de points de son espace d'états. Hopfield détermine de telles conditions (elles ont été depuis très fortement étendues). Lorsqu'elles sont réalisées, on dispose d'un système *autonome,* sans entrée ni sortie distinguées, et qui *peut* cependant servir de système *input/output,* en un sens particulier : un ébranlement initial l'amène dans un nouvel état d'équilibre — l'ébranlement est alors l'*input,* et le nouvel équilibre l'*output.*

Les divergences dans le « camp » connexionniste apparaissent d'entrée de jeu, dès qu'est posée la question du rapport entre les caractérisations matérielle et informationnelle des systèmes cognitifs. Les chercheurs qui se rattachent au courant PDP tendent à adopter le fonctionnalisme classique : ils situent leur modélisation non pas au niveau de la réalisation physique dans le tissu du système nerveux central, mais à celui du traitement de l'information. Que le composant de base soit un neurone formel et l'architecture réticulée facilitera, selon eux, l'implémentation

---

20. Cf. Daniel AMIT, *op. cit. supra* n. 5. « ANN » dénote malheureusement aussi depuis quelque temps *« artificial neural network »,* ce qui tend à faire perdre au sigle sa spécificité.

21. Dans une mémoire informatique classique (en l'absence de dispositifs *ad hoc* d'indexation ou autre), le stimulus ne peut être qu'une « adresse » déterminée une fois pour toute pour chaque item stocké, et la réponse attendue l'information logée à cette adresse. Dans une mémoire adressable par le contenu, le stimulus est une partie du contenu, et la réponse la totalité complétée du contenu mémorisé. La mémoire humaine est généralement rapportée à ce dernier type.

22. John J. HOPFIELD, « Neural Networks and Physical Systems with Emergent Collective Computational Abilities », *Proc. Natl. Acad. Sc. USA,* vol. 79, 1982, p. 2554-2558 ; et *in* J. ANDERSON, E. ROSENFELD, eds, *op. cit. supra* n. 1, p. 460-464.

des réseaux connexionnistes dans des systèmes cérébraux constitués d'assemblées de neurones réels — mais le rapport n'est pas de l'ordre de la ressemblance ou de la simplification[23]. Au contraire, certains connexionnistes considèrent que leur démarche est celle d'une neurophysiologie théorique (ou « computationnelle »), partie intégrante de la biologie théorique ; l'idée d'une séparation radicale de deux niveaux d'explication privilégiés est explicitement rejetée par certains, qui se gaussent du « rêve de Marr »[24]. Plus proches, sur ce point, de ces connexionnistes « biologisants » que de la branche PDP, les connexionnistes de tendance ANN estiment prématuré de se prononcer sur la nature des rapports entre théorie neurobiologique et modèles connexionnistes. La situation ne leur semble pas foncièrement différente de celle qui prévaut dans toute tentative pour attaquer, par les méthodes abstraites de la physique mathématique, un phénomène complexe.

A la distinction bipartite dans les systèmes classiques entre partie fixe et partie variable répond une division tripartite dans les réseaux. La partie vraiment fixe est constituée par l'architecture (nombre des unités et connectivité) et par la loi de transition des unités ; elle correspond en un sens à l'architecture et aux opérations câblées de l'ordinateur non encore programmé, mais elle est sensiblement plus pauvre que la partie fixe du système cognitif classique, qui incorpore les capacités générales de la machine de von Neumann (beaucoup plus complexes que celles du réseau « brut ») et les capacités spécifiques d'un programme. A l'autre extrême, la partie vraiment variable du réseau est constituée par les unités actives au moment considéré ; elle correspond elle aussi à quelque chose dans le modèle classique, à savoir le contenu de la mémoire de travail de l'ordinateur, ou de la mémoire à court terme de maint modèle de la psychologie cognitive. Mais dans le système classique, ces contenus ne sont pas de nature différente des autres représentations, celles, plus stables, de la mémoire à long terme : ensemble, elles sont la partie variable du système, partie beaucoup plus riche, par conséquent, que celle du réseau. Il y a donc, de part et d'autre, un résidu qui dans le réseau se localise dans les connexions — résidu dont le slogan premier

23. Une défense approfondie de ce point de vue est présentée par Paul SMOLENSKY dans un article retentissant, « The Proper Treatment of Connectionism », *The Behavioral and Brain Sciences*, vol. 11, 1988, p. 1-74.

24. David MARR a défendu avec beaucoup de force et de rigueur le principe d'une distinction de niveaux (il en dégage trois) de caractère fonctionnaliste. Voir son célèbre ouvrage posthume, *Vision*, San Francisco, Freeman, 1982 ; ou, pour les passages pertinents, son article dans *Mind Design*, *op. cit. supra* n. 1. Ce sont Patricia CHURCHLAND et Terrence SEJNOWSKI qui s'en prennent à cette conception dans « Neural Representation and Neural Computation », *in* L. NADEL, L. COOPER, P. CULICOVER, R. HARNISH, eds, *Neural Connections and Mental Computation*, Cambridge, MA, MIT Press, 1988.

du connexionnisme dit l'importance : « Toute la connaissance réside dans les connexions », c'est-à-dire dans l'ensemble des poids synaptiques w_{ji}. Comme le programme, les connexions sont fixes au cours d'un processus, mais comme les données représentées, elles sont acquises et reflètent directement des aspects de l'environnement dont la connaissance est indispensable au bon fonctionnement du réseau.

Malgré de notables différences, on discerne donc sur les deux premiers points de doctrine (la distinction entre niveaux et entre parties fixe et variable) une parenté entre les approches, classique et connexionniste. Il en va tout autrement sur les autres points, à savoir le système de représentation, les principes tant physiques que cognitifs régissant l'évolution des systèmes et, enfin, le fondement de la cohérence entre le système et son environnement.

Le système de représentation, tout d'abord, n'est pas un langage formel ; ses « formules » sont des unités, ou des vecteurs d'unités, selon que les représentations sont « localistes » (le support d'une entité sémantique étant une unité du réseau) ou « distribuées » (le support étant une suite ordonnée d'unités, et chaque unité étant inversement impliquée dans la représentation de plusieurs entités)[25]. Dans l'approche PDP, il y a bien une combinatoire des représentations individuelles, mais elle est de type ensembliste, non concaténatoire : les ensembles activés se superposent et s'intersectent — c'est une combinatoire fruste. Dans l'approche ANN, il n'y a pas pour l'heure de combinatoire du tout, mais certains envisagent d'en développer une qui soit fondée sur la théorie des bifurcations[26], ou plus généralement, sur une temporalisation des représentations.

L'évolution d'un réseau obéit non à des calculs figurant des inférences qui s'enchaînent linéairement, mais à un système d'équations différentielles (en général cependant discrétisées) ; c'est un processus de « relaxation », c'est-à-dire de recherche d'une position d'équilibre contrainte par un grand nombre d'interactions simultanées. Sur le plan cognitif, ces interactions correspondent à des associations de force variable, ou encore à des « micro-inférences » non impératives se compensant partiellement, et concourant à un effet global non réductible à une force unique.

N'est-il pas cependant abusif de parler d'inférence, serait-ce à l'abri du préfixe « micro », avant d'avoir précisé de quelle logique il s'agit, et à quoi elle s'applique ? Que serait l'équivalent, dans un réseau, du niveau

25. Cf. D. RUMELHART, J. MCCLELLAND *et al., op. cit. supra* n. 1, vol. 1, chap. 3.

26. Cf. Jean PETITOT, « Hypothèse localiste, modèles morphodynamiques et théories cognitives : remarques sur une note de 1975 », *Semiotica*, vol. 77, 1989, p. 65-119, et son article dans le présent numéro.

syntaxique ? Contentons-nous pour le moment d'observer que la réponse n'a rien d'évident, contrairement à ce que certains écrits connexionnistes voudraient faire croire ; ils emploient, en effet, le terme « syntaxe » au sens de niveau des déterminations physiques commandant les transitions du système. C'est là un abus de langage dangereux, provenant du paradigme classique, dans lequel, on l'a vu, c'est par la médiation de la syntaxe que s'établit le pont entre inférence et cause, ou entre niveau cognitif et niveau physique. Or, dans un réseau, rien ne s'offre immédiatement au regard qui puisse jouer le rôle de syntaxe *indépendamment* des lois de transition[27].

Rien, par conséquent, d'analogue au parallélisme entre syntaxe et sémantique d'un langage formel ne semble pouvoir constituer la garantie du maintien de l'adhésion du système à l'environnement. Mais cette garantie n'est justement pas nécessaire. Dans l'approche PDP, la possibilité d'écarts entre la réponse du système et la bonne réponse est constitutive et censée correspondre à la faillibilité caractéristique des systèmes cognitifs biologiques. On exige toutefois une conformité statistique des résultats obtenus par le réseau, et celle-ci est assurée par l'application d'un principe général de minimisation des écarts[28]. Ce principe a pour effet, étant donné un stimulus s proche d'un exemple s_0 pour lequel le système a préalablement appris la réponse correcte r_0, de pousser le système à fournir, selon les cas, soit une réponse r proche de r_0 et qui en diffère approximativement comme s diffère de s_0, soit tout simplement la réponse r_0. Que le réseau soit capable, par apprentissage au contact de l'environnement, d'ajuster ses paramètres en sorte de pouvoir appliquer le principe général dans cet environnement-là en fait essentiellement un capteur de régularités statistiques ; et qu'il l'applique en complétant un *input* conformément au principe de continuité ou de stabilité qu'on vient d'énoncer en fait essentiellement une machine associative.

Il n'y a donc ici, pour articuler le niveau matériel au niveau cognitif, qu'une seule médiation : les représentations, tant explicites et fugaces (ce sont les unités actives au début du processus), qu'implicites et permanentes (ce sont les poids synaptiques ajustés au cours de l'apprentissage).

---

27. Remarquons qu'il en est de même d'une machine de Turing ou de von Neumann. Contrairement à ce que semblent par moments croire J. Fodor et Z. Pylyshyn, ainsi que maints autres acteurs (qui vont jusqu'à parler du « paradigme de von Neumann », auquel s'opposerait un « paradigme "non-von" »), l'architecture de von Neumann ne suffit pas à imposer la conception classique : celle-ci n'émerge précisément que lorsqu'on fait opérer cette architecture sur la syntaxe d'un langage formel. Lorsqu'on contemple un système classique, il est donc nécessaire de chausser des lunettes « cognitives » (de le considérer comme un système de traitement de l'information) pour y discerner une syntaxe.

28. Ce que P. Smolensky appelle, de manière évocatrice, le principe de maximisation de l'harmonie ; cf. D. RUMELHART, J. MCCLELLAND *et al., op. cit supra* n. 1, vol. 1, chap. 6.

Les premières représentent les entités considérées par le système, et sont traitées non pas conformément à une syntaxe interne, mais au principe d' « harmonie » qui n'est que la description ramassée de la loi de transition du système, déterminée par les traces laissées par l'environnement dans le système sous la forme des poids.

Dans l'approche ANN, on peut envisager de se dispenser complètement de garantie d'adhésion ou de correction. Reprenons l'exemple de la mémoire ; il est caractérisé par l'absence de contrainte sur les réponses : on demande seulement au système de se stabiliser après avoir été exposé à un stimulus significatif. Il est vrai que l'on cherche à obtenir de lui une taxinomie raisonnable : on lui demande de ne pas tout confondre, de ne pas non plus tout distinguer. Mais il n'est plus question alors de correction — seulement de commodité pour l'utilisateur, et le cas échéant, s'il s'agit d'un organisme, pour le système lui-même.

On ne saurait conclure cette présentation des systèmes connexionnistes sans s'interroger sur leur caractérisation comme systèmes « computo-représentationnels ». Ils prétendent, en effet, illustrer une conception de la cognition selon laquelle celle-ci se ramène à des processus calculatoires sur des représentations — et ne se distinguer des systèmes classiques que par la nature des calculs mis en œuvre, ou encore celle des représentations, voire les deux. On dira, par exemple, que les calculs sont parallèles, ce qui les rend foncièrement différents des calculs séquentiels de l'approche classique. Le malheur est que tout calcul au sens d'une manipulation effectuable de signes discrets peut être exécuté de manière séquentielle sur une machine de Turing ou de von Neumann : cela résulte de la célèbre thèse de Turing-Church, mais il n'est pas nécessaire de l'invoquer, puisque les réseaux peuvent (visiblement) être simulés sur des ordinateurs classiques[29], et le sont effectivement. On dira encore que les réseaux manipulent de l'information *numérique* et non *symbolique ;* que la cognition s'explique donc comme un calcul sur des représentations numériques — et non symboliques ! On entend par là que les entités sur lesquelles s'effectuent les calculs ne sont pas les éléments d'un langage ou système formel. Mais toute représentation est par définition symbolique, en sorte que si l'on veut que les nombres représentent, on ne peut les empêcher de le faire exactement comme les symboles classiques. Si, inversement, l'on veut que les calculs des réseaux ne se ramènent pas principiellement à la notion canonique de calcul, il faut les empêcher d'opérer sur des représentations — faire en sorte qu'ils ne soient que l'hypostase descriptive de *processus,* dans le sens où les calculs qui

29. La réciproque est vraie, abstraction faite de la limitation de mémoire.

donnent les trajectoires des planètes ne sont que notre manière de les décrire, les planètes elles-mêmes n'en ayant cure. La suite du présent article apportera peut-être quelques indications utiles sur ce difficile problème, mais il semble dès à présent que *si* les connexionnistes veulent se démarquer à ce très haut niveau de généralité des classiques, il leur faut renoncer soit aux représentations, soit au calcul — sacrifices presque impossibles dans le contexte actuel, puisque aussi bien la notion de représentation est quasiment constitutive des sciences cognitives[30], et que le terme « calculatoire » ou « computationnel » est devenu quasi synonyme de « scientifique » ou « sérieux ».

III. — LA VRAIE NATURE DES PHÉNOMÈNES

L'introspection n'est pas un bon instrument d'investigation scientifique — c'est sur ce rejet[31] que le béhaviorisme prit appui, et les sciences cognitives n'y reviennent pas. Au contraire, bon nombre de leurs succès consistent à mettre au jour des effets invisibles à l'œil introspectif, voire inacceptables pour l'intuition. Cependant, elles réhabilitent les états mentaux — quelque chose, donc, que l'introspection permet de deviner, si partiellement et si trompeusement parfois que ce puisse être. Ce que nous pensons penser, ce que nous croyons croire fournit au moins une indication sur ce que nous pensons et croyons, et notamment sur la manière dont nos pensées et croyances s'enchaînent. Et si l'analyse scientifique ou la simulation révèlent des processus que nous ne parvenons pas à faire coïncider au moins partiellement avec les enchaînements de nos pensées, ce hiatus appelle à son tour une explication.

a. Le connexionnisme sauve les phénomènes

Or, que nous montre le spectacle de nos pensées — ou, si l'on préfère, des événements mentaux dont nous avons conscience ? Prenons d'abord le cas des pensées « rapides » : celles qu'accompagnent la compréhension ou l'émission d'une phrase courante dans notre langue maternelle ;

30. Il existe un courant antireprésentationniste, incarné notamment par J.J. Gibson, W. Freeman et Ch. Skarda, B. Shanon, F. Varela, mais il est très minoritaire (ce qui ne signifie pas, bien entendu, qu'il s'égare nécessairement).

31. Cf., par ex., George SPERLING, « The Magical Number Seven : Information Processing Then and Now », *in* William HIRST, ed., *The Making of Cognitive Science. Essays in Honor of George A. Miller*, Cambridge, Cambridge University Press, 1988 ; William LYONS, *The Disappearance of Introspection*, Cambridge, MA, MIT Press, 1988.

la reconnaissance d'un visage, d'un lieu, d'un objet, d'une voix, d'une situation ou d'un air de musique familier ; la résolution d'un problème simple d'un type parfaitement connu ; une décision de routine ; un déplacement sur le court de tennis pendant un échange de balle ; la traversée en voiture d'un carrefour rencontré des milliers de fois, etc. L'introspection ne nous dit certes pas grand-chose sur ces pensées-là, précisément peut-être parce qu'elles viennent trop vite[32]. Intuitivement, en tout cas, elles se ressemblent, et ressemblent donc toutes aux mieux identifiées d'entre elles, à savoir les perceptions, alors que la tradition philosophique les distingue soigneusement, séparant perception de raisonnement et d'action, et que le cognitivisme veut tout ramener à l'inférence.

Peut-être faut-il donc chercher du côté des pensées plus lentes, celles qui comportent un délai perceptible, indicateur d'un processus complexe nous fournissant une information, une connaissance ou une compréhension qui ne nous sont pas immédiatement accessibles. Examinons donc la manière dont nous parvenons à une conclusion à partir d'informations nouvelles et à la lumière de nos connaissances et de notre expérience ; observons ce qui semble se produire en nous lorsque nous prenons une décision, lorsque nous résolvons un problème, lorsque nous identifions un visage, un objet peu connu, ou encore l'auteur d'un morceau de musique que nous entendons sans doute pour la première fois, lorsque nous lisons un texte difficile, lorsque nous comprenons ce qu'on vient de nous dire dans une langue qui n'est pas la nôtre, lorsque nous nous frayons un passage à travers une foule dense ou pilotons une voiture sur un itinéraire accidenté. Le processus que nous observons en nous-mêmes s'apparente-t-il dans ces cas à une suite d'inférences ? ou bien à la perception d'une scène mal éclairée ? Notre flux mental ressemble-t-il à l'explication que Sherlock Holmes fournit au pauvre Watson après avoir conclu son enquête, ou est-il plus proche de l'*Aha-Erlebnis* que nous vivons au moment précis où, nos yeux s'habituant à la pénombre, nous saisissons comme un tout organisé et chargé de sens les fragments de scène que nous discernions l'instant d'avant ? Sans doute des étapes se sont-elles esquissées fugitivement ; peut-être même avons-nous eu conscience d'essayer d'assembler certaines pièces du puzzle à l'aide d'inférences (« Si ceci est X, alors... », ou « Si je fais d'abord ceci, alors... »). Mais le rôle qu'elles jouent dans la stabilisation finale de nos pensées n'a rien de clair : elles orientent notre recherche (parfois vers une impasse), mais notre impression est généralement qu'elles ne constituent

32. Il s'écoule, par exemple, quelque chose comme un dixième de seconde entre l'apparition d'un visage familier et sa reconnaissance.

pas le tissu même du processus, qui s'est déroulé pour l'essentiel comme une révélation progressive menant à une cristallisation finale.

Des pensées assez lentes aux pensées très lentes, le passage est sans rupture. Nous arrivons dans le domaine du raisonnement mathématique et scientifique, de la délibération, du jeu d'échecs, du diagnostic complexe, de l'expertise. Et nous y trouvons, dressé pour nous depuis longtemps, depuis que les savants, les écrivains, les joueurs d'échecs et les experts nous ont forcés à distinguer le contexte de la découverte de celui de la justification, le même constat : les étapes, les inférences, l'assemblage réfléchi de fragments, jouent un rôle, souvent important, jamais suffisant pour rendre compte du processus. En fait l'expert à son affaire rejoint l'homme ordinaire à la sienne : sa pensée n'est pas lente, elle est au contraire de l'ordre de la perception et du réflexe, comme y insistent Hubert Dreyfus et Stuart Dreyfus[33]. Le grand maître d'échecs voit presque immédiatement ce qu'il va faire, et ne passe son temps qu'à des vérifications, des auto-explications, des comparaisons réfléchies avec les précédents, des spéculations sur les réactions de l'adversaire, sur son état psychique, etc. Ayant en mémoire quelque 50 000 configurations, il reconnaîtrait, selon ces auteurs, celle qu'il affronte en l'assimilant, globalement et sans calcul, à l'une d'elles, et dans le même mouvement choisirait la réaction appropriée.

Admettons la validité de cette description : admettons que, vus ou vécus par nous-mêmes, examinés même par le psychologue réglant son instrument de façon à ne pas perdre complètement ce dont témoigne l'expérience consciente des sujets, nombre de nos processus mentaux tiennent plus de la perception, fût-elle différée, que de l'enchaînement d'inférences. En quoi cela fait-il pencher la balance du côté connexionniste ?

La réponse est évidente, même si elle est, on s'en doute, loin d'être définitive. D'une part, en effet, ce que modélisent, en première analyse, les réseaux vus comme systèmes cognitifs, ce sont des formes, plus ou moins généralisées, de perception (ou, si l'on préfère, songeant notamment à la mémoire associative de Hopfield, de reconnaissance). D'autre part, l'évolution d'un réseau au cours d'une tâche correspond bien aux descriptions que l'on vient de donner des processus mentaux : elle est parfois ponctuée d'étapes — passage d'une hypothèse à la suivante, élimination d'une possibilité, etc. ; elle est parfois orientée par une inférence — si l'on choisit d'interpréter ainsi, par exemple, l'activation de telle sous-population d'unités provoquée par l'activation antérieure de telle autre ; mais elle se déroule pour l'essentiel dans le désordre apparent

33. Cf. *Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer*, New York, The Free Press, 1986, p. 32-35.

des interactions multiples et ne prend un sens clair que lors d'étapes éventuelles et surtout lors de la stabilisation finale. A chaque instant intermédiaire, des hypothèses sont inégalement actives, des liens se font sentir simultanément avec des forces variables, positives ou négatives. Et ces « micro-événements » ne se laissent qu'exceptionnellement sommer en des « macro-inférences ».

D'autres phénomènes sont (en première analyse), sinon sauvés, du moins respectés par le connexionnisme. Il en est ainsi du rôle reconnu depuis les travaux d'Eleanor Rosch aux prototypes dans les tâches de catégorisation[34], où selon les classiques intervenaient des définitions par conditions nécessaires et suffisantes. Il en est ainsi de l'aptitude singulière de l'homme à traiter une information incomplète, incertaine, voire contradictoire : contrairement aux systèmes classiques, les réseaux, comme l'homme, fournissent des réponses raisonnables dans des conditions informationnelles loin de l'optimal, et sans faire intervenir de mécanisme particulier : ce sont les mêmes mécanismes, et les mêmes raisons qui expliquent leur excellent comportement dans de bonnes conditions et leur comportement convenable dans de mauvaises. Il en est ainsi, enfin, de certaines erreurs systématiques chez le sujet normal, et de certains syndromes observés en neuropsychologie : erreurs et déficits humains prennent souvent des formes bien spécifiques que reproduisent les réseaux, le cas échéant lésés artificiellement par exemple par suppression de certaines unités.

Le dernier grand phénomène qui selon certains donnerait un avantage décisif aux réseaux est celui du *contexte.* Il est reconnu par tous que le contexte joue un rôle crucial dans tous les domaines de la cognition, de la communication, notamment verbale, jusqu'au raisonnement, à l'action, voire à la perception. La « sensibilité au contexte » est une qualité prisée chez les humains et poursuivie, tel le Graal, par les spécialistes d'Intelligence Artificielle et plus généralement par les concepteurs de logiciels — c'est la marque même de la véritable intelligence (et son absence la marque de la bêtise, et l'une des principales sources du comique).

Là s'arrête l'unanimité. Il n'est d'abord pas clair que tous les phénomènes auxquels les uns et les autres font référence lorsqu'ils évoquent, ou brandissent, le contexte aient en commun un ensemble significatif de propriétés. Ensuite, tout un spectre de positions se dessine sur l'ampleur de la « contamination » par le contexte : intervient-il aux tout derniers stades des processus, ou le trouve-t-on dès le départ, affectant les primitives mêmes ? Certains stades sont-ils à l'abri, ou bien le contexte intervient-il à tout moment ? Si Fodor, par exemple, défend l'idée de modula-

34. Cf., par ex., Eleanor ROSCH, « Principles of Categorization », *in* E. ROSCH, B. B. LLOYD, eds, *Cognition and Categorization,* Hillsdale, NJ, Erlbaum, 1978.

rité[35], c'est pour protéger certains processus (notamment perceptuels) des excès de la psychologie *« new-look »* qui voit les influences *« top-down »* s'immiscer à tous les stades, comme si le contexte (le *« top »*) pouvait tout faire et tout empêcher au niveau *« down »* : nous faire voir dans la cage du zoo un tigre en l'absence du tigre, ou inversement nous cacher le véritable tigre qui s'apprête, contre toute vraisemblance, à nous dévorer place de l'Opéra.

L'espace logique des positions possibles continue longtemps de se ramifier. Ce n'est pas le lieu de l'explorer exhaustivement. Une partie de l'arbre peut se parcourir rapidement, à partir d'une position classique, de la façon suivante : l'influence du contexte est-elle formalisable ? Si oui, si le contexte est donc constitué d'informations générales complémentaires homogènes aux informations particulières de premier plan, ces informations sont-elles de l'ordre de faits ou de l'ordre de règles ? S'il s'agit de faits, ont-elles pour effet d'enrichir les conclusions (logique monotone) ou de les infléchir (logique non monotone, régime des exceptions) ? S'il s'agit de règles, sont-elles de même niveau que les règles de premier degré, ou d'un niveau supérieur (métarègles) ? Dans ce dernier cas, guident-elles les inférences (par exemple, en imposant un ordre de priorité sur les règles de premier niveau), ou bien les modifient-elles (par exemple, en invalidant certaines règles ou en les modifiant) ? Retournant à la racine de notre arbre, plaçons-nous dans le cas d'une réponse négative : l'influence du contexte n'est pas formalisable. Avant de jeter l'éponge, nous pouvons nous demander si elle est modélisable. Répondre positivement, c'est imaginer soit, du côté des faits, un système de pondération, soit, du côté des règles, un infléchissement du régime inférentiel. (Dans les deux cas, ce qui empêche la modélisation d'être une formalisation est l'absence de règles formelles de pondération ou d'infléchissement[36].) Les prolégomènes pour une logique située, de Jon Barwise[37], et la théorie de la pertinence de Sperber et Wilson[38] sont deux tentatives (inégalement développées) pour rendre raison du contexte en caractérisant son influence sur le régime inférentiel, sans la formaliser au sens étroit du terme.

---

35. Cf. J. FODOR, *op. cit. supra* n. 2.

36. La distinction modélisation/formalisation appelle une analyse plus sérieuse qui nous entraînerait trop loin de notre propos.

37. Cf. « Information and Circumstance », *Notre-Dame Journal of Formal Logic,* vol. 27, *3,* July 1986, p. 324-338, et « Unburdening the Language of Thought », *Mind and Language,* vol. 2, *1,* Spring 1987, p. 82-96. Barwise répond à Fodor, qui l'attaque avec une vivacité signalant l'importance de l'enjeu.

38. Dan SPERBER, Deirdre WILSON, *Relevance : Communication and Cognition,* Oxford, Basil Blackwell, 1986 ; trad. fr. par A. GERSCHENFELD et D. SPERBER, *La Pertinence,* Paris, Minuit, 1989.

Dans la perspective connexionniste, le contexte est en un sens si bien incorporé au processus qu'on peut se demander s'il demeure quelque chose de la distinction entre « texte » et contexte, ou entre premier plan et arrière-plan. Ce qui marque, en effet, la place du contexte dans le fonctionnement du réseau, c'est que les conséquences de l'activation d'une unité ou d'un groupe d'unités U sont fonction de l'état d'activation de l'ensemble R des autres unités : le fait, la situation, l'événement ou l'entité e que représente U est donc interprété(e) différemment selon que son environnement v, représenté par R, est dans un état plutôt qu'un autre. Mais il n'existe en vérité aucune dissymétrie intrinsèque entre U et R, et l'on pourrait dire aussi bien que le réseau traite v dans le contexte de e. D'autre part, l'influence du contexte est peut-être ainsi figurée, elle n'est ni expliquée ni contrôlée. Enfin, une pareille conception du contexte tombe sous le coup des critiques adressées très tôt à l'Intelligence Artificielle classique[39] : elle limite préalablement le contexte à un nombre fini de situations identifiées — l'essence de la notion ne se trouve-t-elle pas ainsi niée ? On peut en discuter, mais le connexionnisme ne semble pas avoir sur ce point d'avantage décisif sur le cognitivisme.

Une proposition plus spécifique et plus originale a été faite par P. Smolensky[40] : la sémantique des unités est elle-même dépendante du contexte, dans le cas des représentations distribuées. Le café n'est pas tout à fait la même chose lorsqu'il remplit une tasse et lorsqu'il tache une chemise — ce sont des sous-populations légèrement différentes d'unités qui représentent « café » dans le contexte de la tasse et dans celui de la chemise. C'est pour le coup à un pilier du classicisme que l'on s'attaque ici : les primitives sémantiques sont classiquement indépendantes du contexte, et la sémantique est compositionnelle, ce qui signifie que les sens complexes sont exactement obtenus par des combinaisons appropriées de leurs composants, et elles-mêmes indépendantes du contexte. Il y aurait donc là une possibilité nouvelle — encore faut-il qu'elle reçoive un commencement de concrétisation.

b. Le connexionnisme perd le phénomène central

J. Fodor et Z. Pylyshyn[41] estiment que le connexionnisme est fondamentalement inadéquat comme théorie de la cognition, pour la raison

39. Cf. Hubert L. DREYFUS, *What Computers Can't Do. The Limits of Artificial Intelligence,* New York, Harper, 1972, 2d ed., 1979 ; trad. fr. par Rose-Marie VASSALLO-VILLANEAU, *Intelligence artificielle : mythes et limites,* Paris, Flammarion, 1984.

40. Reprenant une suggestion de Z. Pylyshyn ! *Cf. op. cit. supra* n. 23.

41. *Op. cit. supra* n. 3, et J. A. FODOR, *Psychosemantics. The Problem of Meaning in the Philosophy of Mind,* Cambridge, MA, MIT Press, 1987, Appendice : « Why there still has to be a language of thought ».

qu'il ne fait pas place aux représentations structurées, lesquelles sont seules susceptibles d'expliquer un aspect central de la cognition. Essayons de comprendre ce qu'ils veulent dire.

Voyons d'abord de quel aspect il s'agit. Le langage est notoirement *productif* et *systématique* : il permet d'engendrer une infinité de phrases à partir d'un nombre fini de mots et de règles ; et ses phrases se distribuent sur tout un espace logique, sans laisser d'interstice, au sens où dès qu'y figure quelque chose comme « Jean aime Marie », y figurent nécessairement aussi « Marie aime Jean », « Chacun aime Pierre », etc., et dès qu'y figure « La vache est brune et le cheval est gris », y figurent « La vache est brune » et « Le cheval est gris ». Maîtriser une langue, c'est notamment être en mesure de comprendre et d'émettre une infinité potentielle de phrases (productivité), et de ne pouvoir comprendre ni émettre « Jean aime Marie » sans pouvoir comprendre et émettre « Marie aime Jean », etc. (Maîtriser une langue, c'est donc, explique lumineusement Fodor, tout différent de connaître par cœur une liste de phrases telles que les fournissent les guides pour touristes étrangers.) Or, dit Fodor, ce qui est vrai du langage et de nos capacités linguistiques est vrai de nos pensées et de nos capacités cognitives. La raison ? Celle d'abord que le langage exprime la pensée — comprendre une phrase, c'est saisir la pensée que cette phrase exprime. Mais une réflexion directe sur les pensées dont nous sommes capables conduit à la même conclusion : si nous pouvons penser qu'il fait froid, nous pouvons nécessairement penser qu'il fait très froid, qu'il fait très très froid, etc. ; si nous pouvons penser que le chat est sur le tapis, nous devons pouvoir penser que le tapis est sur le chat ; si nous pouvons vouloir lever le pied gauche et le bras droit, nous devons pouvoir vouloir lever le pied gauche et vouloir lever le bras droit ; etc. (On remarquera en passant la tranquillité avec laquelle Fodor s'appuie sur l'introspection la plus élémentaire, sans même chercher une caution — il faut lui en savoir gré — dans d'hypothétiques expériences de psychologie.)

Voyons maintenant comment Fodor passe de ce constat à l'hypothèse du rôle fondamental des représentations structurées dans la cognition. Depuis Frege, nous rappelle-t-il, nous disposons d'une excellente explication de la productivité et de la systématicité du langage : les formes linguistiques sont obtenues par combinaison de formes primitives ; elles sont structurées en ce sens bien précis, et les processus linguistiques, à tous les niveaux, sont sensibles à la structure de ces représentations. D'un autre côté, on ne dispose d'aucune autre explication de ces phénomènes. La seule attitude rationnelle devant la manifestation des mêmes phénomènes dans un autre domaine, celui des pensées et des processus mentaux, est d'adopter le même genre d'explication. A défaut, donc, d'une

meilleure idée, on mettra au cœur des *états* mentaux des représentations structurées (au sens de la linguistique) — ce seront des formules dans un langage formel interne, le « langage de la pensée » ou « mentalais » ; et l'on attribuera aux *processus* mentaux la propriété essentielle d'être sensibles à la structure des représentations constitutives des états sur lesquels ils opèrent.

Dernière étape, la réfutation du connexionnisme. Les réseaux sont certes des machines à manipuler des représentations, nous disent Fodor et Pylyshyn. Mais leurs représentations ne sont pas structurées. Et à supposer qu'on puisse leur attribuer une structure, les manipulations ne pourraient y être sensibles, car elles reposent sur un principe mathématico-physique de minimisation d'écart, et sur un principe psychologique ou cognitif d'association selon des régularités statistiques — rien qui puisse, sinon accidentellement, assurer une sensibilité à la structure composite des formes matérielles (les populations d'unités activées) répondant à la structure syntaxique des entités représentées.

IV. — LES NIVEAUX

La raison évidente pour laquelle aucun des arguments exposés à l'instant ne saurait par principe être décisif n'est ni leur imprécision, ni leur source introspective ; elle réside dans la possibilité qu'a toute théorie — classique, connexionniste ou autre — de faire appel à une distinction de niveaux. Les classiques, mis en cause par les vertus « phénoménologiques » du connexionnisme, peuvent prendre appui sur l'idée qu'une véritable théorie des phénomènes que celui-ci ne fait somme toute que singer fera nécessairement appel à des entités et à des processus fondamentaux se situant à un niveau inférieur, niveau dont émergeront les propriétés phénoménales en question. Inversement, les connexionnistes, mis au défi par l'objection des représentations structurées, peuvent espérer produire celles-ci à un niveau supérieur, comme effet émergent des processus fondamentaux dont le connexionnisme énonce les lois.

L'appel aux distinctions de niveaux n'est pas l'apanage des théories de la cognition, bien au contraire. Mais celles-ci ont la particularité d'avoir lié leur sort dès l'origine à la postulation d'un niveau particulier, sur la nature duquel les doutes continuent de planer. Les entités qui peuplent ce niveau ne sont pas seulement inaccessibles à l'observation ; elles ne se rangent même pas clairement dans l'une des deux catégories d'inobservables distinguées, comme le rappelle souvent Dennett, par Reichen-

bach[42] : sont-ce des *illata,* c'est-à-dire des objets dont on postule qu'ils existent véritablement, à titre d'entités matérielles — comme des électrons ou des mésons ou des trous noirs ? ou sont-ce des *abstracta,* c'est-à-dire des entités purement théoriques, utiles, voire indispensables, pour ériger des théories portant sur certains objets matériels, mais ne faisant pas elles-mêmes partie de ces objets — comme des centres de gravité ou des trajectoires ?

Les entités dont les sciences cognitives parlent — l'information, les représentations, le contenu sémantique, le traitement de l'information... — demeurent profondément mystérieuses, ce qui n'empêche nullement bon nombre de programmes de recherche de progresser, mais qui rend difficile tout débat sur les fondements. Comme le dit Barwise[43], notre situation rappelle celle des hommes de l'âge du bronze, qui maniaient le bronze fort bien sans posséder ce que nous considérerions aujourd'hui comme une théorie acceptable du bronze.

Le statut ontologique et épistémologique du niveau informationnel, ou représentationnel, est pour les classiques d'obédience fodorienne un dogme — un dogme moderne sans doute : chacun reste libre de le rejeter. Mais, nous avertit charitablement Fodor, celui qui choisit de le faire s'exclut par là même de la science cognitive ; il fait de la biologie, de l'électronique, de la physique, ce qu'on voudra, il n'est plus dans la course, pas plus que ne demeurerait géologue celui qui prétendrait réfuter les théories régnantes en géologie en ne parlant que de la théorie quantique des champs. Ce n'est pas la possibilité d'une science cognitive que Fodor pose comme dogme ; c'est l'existence du niveau représentationnel comme condition nécessaire de cette possibilité. Et si le connexionnisme est à ses yeux la nouvelle Carthage à détruire, c'est qu'il prétend fonder la science cognitive sur un autre niveau, et en faire la science de ce niveau.

Mais quelle est la nature de la hiérarchie même qu'implique la référence à un niveau privilégié ? Voilà qui n'est pas clair non plus, et qui mériterait à soi seul de longs développements. Parle-t-on de niveaux de description ? Parle-t-on d'échelles de grandeur et de niveaux d'organisation, chacun étant peuplé d'agrégats d'entités peuplant le niveau immédiatement inférieur ? Parle-t-on de niveaux d'intégration, comme les neurophysiologistes ?

Les classiques parlent essentiellement de niveaux de description, et en

---

42. Hans REICHENBACH, *Experience and Prediction,* Chicago, University of Chicago Press, 1938, p. 211-212 ; cité *in* Daniel C. DENNETT, *The Intentional Stance,* Cambridge, MA, MIT Press, 1987, p. 53, trad. fr. à paraître, Paris, Gallimard.

43. « Information and Circumstance », *art. cit. supra* n. 37.

même temps, secondairement, de niveaux d'abstraction : le niveau privilégié est donc obtenu par une double opération, l'une, cruciale et originale à leurs yeux, de visée, l'autre, secondaire et banale, d'idéalisation. Acceptons sans la discuter ici la seconde, et décrivons la première. Les états cérébraux d'un organisme sont classés selon la relation d'équivalence « joue le même rôle que... », le rôle étant celui que tient un état mental vis-à-vis des autres états internes et des états perceptuels et moteurs. Mais bien entendu, c'est sous une certaine description que ces états peuvent être vus comme interagissant les uns sur les autres (sont-ce des molécules de bois, ou bien une latte de parquet posée de guingois qui provoquent ma chute ? ou bien encore ce qu'elles provoquent serait-il la modification de la trajectoire des atomes qui me composent ?). On ne peut donc que *postuler* que ces états sont notamment informationnels, ou comme on dit parfois, « sémantiquement évaluables » : le fonctionnalisme offre une solution au rapport corps-esprit, mais à un coût élevé ; en ce sens, il n'est aucunement réductionniste. Le niveau informationnel demeure infondé. Admettons donc qu'un état mental soit un état cérébral caractérisé sémantiquement ou informationnellement — ou plus exactement la classe d'équivalence des états cérébraux ayant le même rôle informationnel dans l'économie du système. Que seront alors les *processus* mentaux ? C'est ici que la métaphore de l'ordinateur joue un rôle crucial : l'équivalent des états mentaux sont les états internes de la machine de Turing (ces états sont eux-mêmes, et c'est heureux, des abstractions ; ce sont des classes d'équivalence dans un espace abstrait de configurations) ; or ces états se transforment les uns dans les autres sous l'effet des opérations de la machine, opérations parfaitement définies sur le plan abstrait, et dont la réalisation matérielle est maîtrisée théoriquement : il n'y a pas de mystère dans la façon dont un calculateur concret réalise ou incarne les opérations d'une machine de Turing. *Ergo* les opérations qui font passer d'un état mental à un autre sont identiques ou du moins semblables aux opérations d'une machine de Turing, et il n'y a pas de mystère *théorique* dans l'idée que les processus cérébraux réalisent ces opérations. (Il y a certes un mystère empirique, dont la solution est d'autant plus impatiemment attendue qu'elle devra révéler le pourquoi de la stabilité de ces processus vis-à-vis de la relation d'équivalence sur les états neurophysiologiques : il faut que deux états équivalents soient transformés en deux autres états équivalents, ce qui est une contrainte extrêmement forte.)

Comment concevoir alors le rapport entre le niveau qui est celui des classes d'équivalence d'états cérébraux informationnels et le niveau immédiatement inférieur ou sous-jacent ? La question est ambiguë, et c'est là une source de confusions. *Primo*, en tant que niveau de description, il a pour soubassement le niveau des états cérébraux eux-mêmes

(une classe d'états cérébraux renvoie, dans ce rapport, à n'importe lequel de ses éléments), mais bien entendu *sous une certaine description* ; cette description est celle que fournissent les neurosciences. Mais les neurosciences fournissent *plusieurs* descriptions, selon le niveau d'*organisation* auquel elles se placent. De laquelle s'agit-il ? Nécessairement de celle qui porte sur des entités de la « taille » de celles qui peuplent le niveau informationnel : si ce sont, par exemple, des assemblées de neurones d'un certain type qui sont porteuses de représentations atomiques variables, des colonnes de neurones d'un autre type qui sont porteuses de représentations logiques fixes, le niveau cherché est celui de ces assemblées et de ces colonnes. *Secundo,* ce niveau est à son tour en relation avec un niveau d'organisation immédiatement inférieur, qui pourrait être par exemple celui des neurones individuels caractérisés par le traitement de certains signaux qu'ils effectuent.

Finalement le rapport dont parlent les classiques, qui est celui de l' « implémentation » du niveau informationnel dans le « wetware », est purement conceptuel, vide de contenu empirique — on conçoit mal une preuve empirique de la non-existence d'un niveau d'intégration neurophysiologique correspondant au niveau informationnel, ou de la non-existence de processus neurophysiologiques réalisant les opérations postulées à ce dernier niveau, et stables pour la relation d'équivalence. Il n'y a, dans la doctrine classique, aucune théorie substantielle de l'émergence d'entités sémantiques à partir d'un niveau fondamental — contrairement à ce qui est parfois suggéré, le candidat avancé étant l'implémentation ou instantiation. Au contraire, l'existence de niveaux d'intégration du tissu nerveux menant par paliers du neurone individuel à un niveau « sous-jacent » au niveau informationnel est riche de contenu empirique ; en ce sens, il est vrai que les sciences cognitives peuvent avoir pour les neurosciences une valeur heuristique, alors qu'inversement celles-ci restent sans aucune influence sur une conception *classique* rigoriste des sciences cognitives. Lorsqu'on parle des contraintes qu'exerceraient les uns sur les autres les différents niveaux de description, et donc les différentes sciences du « cerveau/esprit », on va trop vite.

Le tableau classique se complique encore par le fait que le niveau informationnel lui-même peut se ramifier indéfiniment, on l'a vu, par le jeu de machines virtuelles nichées les unes dans les autres. Si bien que souvent, lorsqu'il est question de la relation classique entre niveaux, on parle d'une relation exacte d'implémentation — il vaudrait mieux dire de compilation — entre deux langages ou deux machines virtuelles. Si l'on veut par exemple, dans le cadre classique, faire droit à l'idée d'un niveau de micro-entités et de micro-processus, sous-jacent au niveau des phénomènes mentaux « visibles », par exemple accessibles à l'introspection, on

postulera un langage « machine » ou de niveau inférieur, considéré comme définissant les « véritables » opérations physiques de la machine, et un langage « évolué », « compilé » ou de niveau supérieur, dans lequel s'inscrivent, ou au niveau duquel émergent, les « macro »-phénomènes. Mais, ajoutera-t-on, il existe par définition une réduction sans résidu du niveau supérieur au niveau inférieur : les opérations « réelles » rendent *exactement* compte des phénomènes émergents ou virtuels commodément décrits dans le langage évolué.

C'est précisément à cette réduction sans résidu que veut échapper le connexionnisme. Le sous-titre même de la bible du courant PDP, « investigations de la *micro-structure* de la cognition », indique suffisamment son ambition de découvrir le niveau fondamental auquel les choses se passent réellement, c'est-à-dire auquel on peut donner une description systématique, exhaustive et exacte des *processus* en jeu dans la cognition. Or quel est ce niveau ? Il ne peut s'agir « du » niveau physique, c'est-à-dire d'un niveau de description physique ou biologique des états cérébraux. Il importe, en effet, de « rester dans la course », c'est-à-dire de parler d'entités sémantiquement évaluables, d'états et de processus informationnels. Sur ce point, les connexionnistes de tendance PDP, comme on l'a vu, acceptent la règle de Fodor : ils sont fonctionnalistes, et refusent de se laisser cantonner, comme les y invitent Fodor et Pylyshyn[44], au rôle d' « implémentateurs ». Ils refusent, en d'autres termes, d'être ceux qui montrent comment on peut réaliser les fonctions classiques dans un substrat (abstrait) — on parle dans ce contexte d' « architecture » — non classique.

Il leur faut donc postuler l'existence d'un niveau plus fondamental ou « micro » que le niveau « symbolique » des classiques, mais encore (dans l'image d'une descente de la cognition vers la matière, qui mène en même temps des phénomènes les plus complexes aux moins complexes) informationnel. Cependant, s'ils veulent que ce niveau soit effectivement différent du niveau symbolique (ce n'est pas le cas de tous les connexionnistes : beaucoup se contentent de proposer une théorie rivale au même niveau que les classiques), ils doivent éviter de le munir d'une sémantique classique. Ils postulent donc un « saut » *(shift)* sémantique[45] entre les macro-entités, les assemblées d'unités, et les micro-entités que sont les unités individuelles. Celles-ci, dans le paradigme des représentations distribuées que l'on a mentionné au § 2, se voient munies de particules de sens dont seul l'assemblage en cours de traitement peut fournir un véritable sens ou une représentation au sens classique. Ainsi, le niveau

44. « Connectionism as a Theory of Implementation » est le titre d'une section de leur *op. cit. supra* n. 3, p. 64-66.

45. Cf. P. SMOLENSKY, *art. cit. supra* n. 23.

fondamental serait-il « subsymbolique » ; il serait caractérisé par une « syntaxe » exacte, c'est-à-dire par des transformations exactement spécifiables dans le vocabulaire de la physique mathématique, et par une « proto-sémantique » floue. Au niveau supérieur d'organisation émergeraient des phénomènes affectant des assemblées d'unités munies d'une pleine sémantique, mais régies par une « syntaxe » floue : toute caractérisation des transformations les affectant serait approximative. La description au niveau supérieur ne se réduirait donc pas à la description au niveau inférieur, et le rapport entre les niveaux se rapprocherait plus de celui qui s'établit entre microphysique et physique newtonienne que du rapport de compilation entre langages informatiques.

Remarquons en passant que sous l'angle de la sémantique et de la « syntaxe » (toujours dans le sens, abusif, de système des lois de transition entre états), l'existence de sauts crée donc un espace logique à quatre positions, toutes occupées, présenté sans autre explication dans le tableau suivant :

Saut sémantique Saut « syntaxique »	NON	OUI
NON	Psychologie cognitive classique des processus « personnels » ; Intelligence Artificielle classique	Psychologie cognitive classique des processus « subpersonnels » ; p. ex. psycholinguistique, vision artificielle
OUI	Connexionnisme modéré de type « localiste » ; psychologie du sens commun *(folk psychology)*	Connexionnisme radical ; paradigme subsymbolique

Pour séduisante qu'elle soit par certains aspects, la position connexionniste radicale défendue par Smolensky (le « paradigme subsymbolique » occupant la case la plus « exotique » du tableau) demeure fragile. Car en refusant d'assimiler le niveau « subsymbolique » au niveau informationnel des classiques, elle se prive d'un solide ancrage dans l'intuition et l'introspection : que nos états mentaux soient sémantiquement évaluables, ou si l'on préfère intentionnels au sens de Brentano — qu'ils visent quelque chose qui leur est extérieur — est certes un grand mystère, mais porte en même temps la marque de l'évidence. Les entités subsymboliques seraient intentionnelles elles aussi, mais sans qu'on puisse spécifier dans le vocabulaire de la psychologie, même étendu, à quoi elles ren-

voient. Et en refusant, d'autre part, d'assigner au niveau subsymbolique, serait-ce au prix d'une schématisation, une place parmi les niveaux d'intégration ou d'organisation du tissu cérébral, ce connexionnisme-là se prive d'un solide ancrage dans la tradition de la modélisation physique. Dire, comme le fait Smolensky, que ce niveau est intermédiaire entre celui des classiques et celui des neurosciences n'est guère défendable — ce n'est sans doute pas même cohérent, si du moins l'on accepte l'analyse qui vient d'être proposée du niveau fondamental des classiques, dont Smolensky entend conserver la conception fonctionnaliste. Les connexionnistes qui interprètent leurs propres efforts comme des prolégomènes à une neurophysiologie théorique ou « computationnelle » (je préférerais, tout simplement, « mathématique ») se placent sur un terrain plus solide, même s'ils doivent faire face à de difficiles objections concernant la « plausibilité » neurophysiologique de leurs modèles.

V. — LES STRUCTURES, LE TEMPS ET LA SIGNIFICATION

Le reproche fondamental adressé par Fodor et Pylyshyn aux réseaux, qui est d'être constitutionnellement incapables de recevoir et de manipuler adéquatement des représentations structurées, n'a pas laissé les connexionnistes indifférents. Ils ont de fait proposé plusieurs procédés de représentations des structures dans les réseaux ; la première proposition, due à Geoffrey Hinton[46], précède même de plusieurs années l'interpellation par les défenseurs du classicisme !

S'il est impossible dans le cadre du présent article de passer en revue les solutions avancées, on peut indiquer certaines des conclusions qui se dégagent d'un tel examen. En premier lieu, il n'existe aucun système couvrant tous les aspects de la représentation et de la manipulation des structures. Il est même assez difficile de discerner, dès que l'on quitte le cadre classique, avec ses formules d'un langage formel, ce qu'ont en commun les procédés connexionnistes de représentations de relations (« Pierre est le père de Marie », « Marie est la sœur de Paul »...), ou de suites finies ($a_1 b_2$ c_3...), de parcours d'une liste, de représentations de classifications (les A sont des B ou des C, les B des X, des Y ou des Z, les C des U, des V ou des W, etc.), de réalisation de « systèmes de production » (au sens de Newell : systèmes modifiant graduellement une base de données par l'application de « règles de production » de la

46. « Implementing Semantic Networks in Parallel Hardware », *in* G. HINTON, J. ANDERSON, eds, *op. cit. supra* n. 5.

forme : si A(x) et B(x), alors C(x) — si, pour une valeur a, figurent dans la base à un instant donné les formules A(a) et B(a), on ajoute à la base, à l'instant suivant, C(a)). En second lieu, on peut distinguer trois types de solutions : celles qui n'affichent d'autre ambition que d' « implémenter » (en un sens qui demande à être précisé) des langages ou des traitements classiques ; celles qui consistent à mettre en évidence, par exemple par une analyse statistique de l'activité des unités au cours du traitement des différentes données, une hiérarchisation de ces unités reflétant la classification naturelle des données ; celles, enfin, qui consistent à inscrire dans le medium des unités des suites de symboles, sans perdre l'esprit anticlassique de la modélisation connexionniste[47].

De ces trois approches, la plus intéressante sur le plan théorique est évidemment la troisième. Mais les solutions auxquelles elle a conduit jusqu'à présent sont à la fois très compliquées et très en deçà de nos attentes car, d'une part, elles limitent *a priori* la longueur ou la complexité des représentations et, d'autre part, elles ne ménagent pas un libre recours à la récurrence. Quoi qu'il en soit, il est intéressant de repérer les différences par rapport à la solution classique. La première est que le réseau doit se subdiviser de manière permanente en sous-réseaux, chaque sous-réseau étant dédié à la représentation particulière d'un rôle ou place dans la structure[48]. La seconde est que pour manipuler les représentations, le réseau doit faire appel au temps d'une manière beaucoup plus fondamentale qu'un système classique. En effet, pour exécuter la transformation de XYZ en X'Y'Z', le réseau doit prendre à l'instant t la configuration XYZ, et à l'instant suivant t' la configuration X'Y'Z' : ne pouvant « dire » XYZ, comme le système classique, le réseau « est » en un sens, ou encore « mime » XYZ, donc pour « être » ou « mimer » X'Y'Z', il doit se transformer — quitte du reste à perdre la trace de son état antérieur et à oublier ainsi la provenance de son nouvel état. Le système classique, lui, dispose d'un tableau noir sur lequel il peut inscrire XYZ, et aussi, sans même l'effacer, X'Y'Z' : l'essence d'un tel système est la division intangible entre la partie variable, lieu des inscriptions, et la partie fixe, transformateur des inscriptions.

On peut donc se demander si le connexionnisme ne doit pas profiter

---

47. Trois exemples caractéristiques de ces approches sont respectivement : 1. David S. TOURETSKY, « BoltzCONS : Dynamic Structures in a Connectionist Network », CMU-CS-89-182 Technical Report, Carnegie Mellon University, August 1989 ; 2. Jeffrey ELMAN, « Representation and Structure in Connectionist Models », CRL Technical Report 8903, Center for Research in Language, University of California at San Diego, 1988 ; 3. P. SMOLENSKY, « Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Networks », *Artificial Intelligence,* sous presse.

48. Je simplifie : le schéma imaginé par Smolensky (cf. *supra* n. 47) est plus subtil ; il implique néanmoins une partition fixe du réseau en sous-réseaux spécialisés.

de sa différence pour opérer un changement radical de perspective, plutôt que de se contorsionner pour imiter le classicisme : s'il le fait trop fidèlement, il risque, en effet, de n'être qu'implémentation (ce qui peut comporter d'intéressantes retombées techniques : il y a là sans doute pour l'Intelligence Artificielle le moyen de surmonter certaines de ses faiblesses) ; s'il ne parvient que partiellement et maladroitement, c'est l'intérêt même de la tentative qui est en cause. Or, que suggère l'insistance des connexionnistes sur leur intérêt pour la « microstructure » de la cognition ? C'est la distinction simple dans son principe, mais si difficile à énoncer clairement dans le cas de la cognition, entre l'*organe* et ses *produits* (qui n'est autre, on peut le présumer, que la vénérable mais plus obscure distinction entre la structure et la fonction). Cette distinction est gommée dans la perspective classique, mais fait retour sous la forme du fantôme dans la machine — celui pour lequel les représentations, si commodément inscrites sur le tableau noir intérieur, représentent. L'intérieur reflète l'extérieur, le langage de la pensée (producteur, agent) mime le langage tout court (produit, agi), comme dans l'univers de la Renaissance — « univers de miroirs dans lequel tout se reflète dans tout » (selon l'expression de R. Gadoffre).

Pourquoi le connexionnisme ne tenterait-il pas de concevoir une machine dont le fantôme serait d'emblée exorcisé, au lieu d'attendre de l'être à la fin, combien hypothétique, des travaux, et qui aurait sa propre notion de ce qui représente quelque chose, de ce qui *pour elle* possède une signification ? Une machine que l'on doterait progressivement des capacités cognitives fondamentales ? La première de ces capacités serait la mémoire d'objets isolés — objectif qui, convenablement généralisé, peut être considéré comme celui de la première vague du connexionnisme contemporain. La seconde capacité fondamentale viserait, par un nouveau biais, le problème des structures ; il s'agit de la mémoire de suites enchaînées d'objets. Ces suites jouent, en effet, un rôle fondamental : toute la cognition, tout le comportement intelligent reposent sur la reconnaissance et la production de suites : suite de notes (air de musique), suite de phonèmes, de mots, de phrases, suite d'événements identiques (coups de cloche), d'où la numération, les suites de déductions, les suites de gestes... Pour obtenir, à partir de là, les représentations structurées des classiques, il ne manquerait que la capacité de typer certains objets.

C'est précisément à ce programme que s'est attelé, avec succès, un connexionniste physicien d'obédience ANN, Daniel Amit[49]. Souvenons-nous que les réseaux, étant munis de connexions multidirectionnelles

49. Cf. *op. cit. supra* n. 5, et « Neural Networks Counting Chimes », *Proc. Natl. Acad. Sc. USA,* vol. 85, 1988, p. 2141 sq.

(donc de *feedback*), sont ici vus non comme des machines entrée/sortie, mais comme des systèmes dynamiques doués d'une dynamique endogène, sujette à des perturbations provenant de l'environnement. Tout retour rapide à l'équilibre est, par définition, un événement cognitivement significatif pour le réseau — ce qui n'est pas arbitraire, dans la mesure où un tel événement est repérable par un certain type de neurone formel situé à l'extérieur du réseau. L'attracteur vers lequel le réseau vient de converger constitue le contenu de l'événement ; et ce contenu est à son tour la représentation de l'événement perturbateur initial. On pourra dire, revenant maintenant à l'idée de Hopfield[50], que le réseau a le souvenir *de* cet événement, souvenir dont le *contenu* est l'attracteur auquel il a conduit le réseau. Arrêtons-nous un bref instant sur cette proposition : quel saisissant contraste par rapport à la doctrine classique, et même par rapport au connexionnisme PDP ! Rien d'équivalent à ce critère de signification intrinsèque (de « signifiance ») ne semble se présenter de façon naturelle dans ces derniers. Tout stimulus ne doit-il pas être significatif, puisqu'il ne peut qu'ébranler un système qui, en l'absence de stimulus, est au repos ? On pourrait être tenté de restreindre la qualité de signifiance aux stimulus conduisant le système à l'équilibre — mais c'est là une propriété indécidable : on le sait dans le cas classique (c'est le célèbre problème de l'arrêt de Turing), et c'est encore vrai, pour de toutes autres raisons, dans le cas des réseaux PDP. Ce qui, en pratique, nous ramène au non-critère précédent. On pourrait aussi délimiter arbitrairement le domaine de la signifiance, ce qui reviendrait, quelle que soit la solution technique adoptée, à faire attribuer au système une valeur distinguée (« non significatif ») à certains stimuli. Mais qui ne voit que *pour le système* un stimulus ainsi traité serait bel et bien significatif ?

Venons-en à la seconde contribution d'Amit. Le problème initial, posé par Donald Hebb[51], était le suivant : quoique deux coups de cloche soient, en tant que stimuli, indiscernables, comment se fait-il que nous n'y réagissions pas de la même façon — que par exemple nous disions « 1 » après l'un, « 2 » après le suivant ? La solution d'Amit prend la forme d'un réseau capable de compter — dans les deux sens du terme : il sait compter « en l'air », c'est-à-dire réciter (un segment initial de) la suite des entiers, et il sait compter les éléments d'une suite d'événements identiques. Voici comment. Le réseau est muni de deux types de connexions, les synapses rapides, qui ont les caractéristiques habituelles,

50. Cf. *op. cit. supra* n. 22. C'est par commodité que j'attribue l'idée à cet auteur : il en partage la paternité avec plusieurs autres chercheurs (T. Kohonen, S. Grossberg, etc.), dont certains revendiquent même l'antériorité.

51. Donald O. HEBB, *Essay on Mind*, Hillsdale, NJ, Erlbaum, 1980.

et les synapses lentes, qui ne transmettent une quantité non négligeable d'influx qu'après un certain temps d'accumulation. La dynamique du réseau peut être vue, en première approximation, comme résultant de la superposition d'une dynamique rapide et d'une dynamique lente. Partant d'un état initial quelconque, le réseau atteint rapidement, soit spontanément (comptage « en l'air »), soit suite à un premier coup de cloche, une position d'équilibre provisoire (« quasi attracteur » de sa dynamique rapide). Au bout d'un certain intervalle de temps, dont l'ordre de grandeur est déterminé précisément (il est compris entre celui du retour à l'équilibre selon la dynamique rapide, et celui de la stabilisation « définitive »), la dynamique lente, aidée, dans le cas des coups de cloche, par un nouveau coup, déstabilise le système, qui regagne rapidement une nouvelle position d'équilibre provisoire, et ainsi de suite jusqu'à la stabilisation finale. Pendant ce temps, un neurone spécialisé détecte les phases successives de convergence vers un quasi-attracteur, et fait « tourner » un compteur. Il détecte également la stabilisation finale, ce qui permet au compteur de se remettre à zéro (pour éviter de compter « douze » au premier coup de minuit...).

Ce réseau n'est pas seulement doté de la capacité de compter des événements cognitivement significatifs ; il possède aussi une notion de temps intrinsèque, donnée par les stabilisations successives dans des quasi-attracteurs. Ce temps n'est pas celui de la pulsation des mises à jour des activités, temps uniforme dépourvu de signification pour le système, temps des transitions « inconscientes » entre deux états « conscients » ; ce n'est pas davantage le temps de l'horloge interne de l'ordinateur classique, également dépourvu de signification, incapable de distinguer, parmi les événements qui le scandent, ceux qui seraient significatifs. C'est vraiment le temps *pour* le système, et non pas seulement le temps *du* système. Voilà donc un concept — certes très particulier —, le temps, représenté *dans* et *pour* le système ; non par un symbole arbitraire, mais par une unité liée de façon très spécifique à l'ensemble du réseau (le neurone détectant les retours rapides au quasi-équilibre).

***

Le connexionnisme — peut-être serait-il préférable de parler des connexionnismes — ne fournissent pas, c'est bien clair, toutes les réponses ; encore n'avons-nous pas évoqué toutes leurs difficultés (notamment celles que pose le passage à des problèmes en vraie grandeur : seuls pourraient les résoudre des réseaux beaucoup plus grands que ceux que l'on utilise aujourd'hui ; or de tels réseaux sont parfaitement incontrôlables dans l'état actuel de nos connaissances — c'est la

version connexionniste de l'explosion combinatoire). Mais le plus difficile, pour le théoricien, est encore de déterminer les questions auxquelles ils apportent des réponses. C'est du reste un trait qu'ils partagent avec le classicisme (au sein duquel il faudrait aussi, mais c'est une autre histoire, distinguer plusieurs doctrines et un grand nombre de programmes de recherche). Chacune de ces approches apporte des solutions dont on a le plus grand mal à préciser la place exacte dans l'économie générale de la cognition. Le classicisme permet, semble-t-il, de comprendre comment des représentations élémentaires peuvent être combinées de manière effective pour permettre à un système ou agent cognitif de déployer des comportements complexes adaptés ou intelligents. Le connexionnisme PDP nous donne une conception étendue de la perception comme détection dans l'environnement de régularités statistiques d'ordre arbitrairement élevé. Le connexionnisme ANN nous aide à concevoir des systèmes auto-organisés, mus par une dynamique interne, capables de distinguer dans le flux des processus certains événements significatifs, munis de concepts intrinsèquement significatifs.

Tantôt ces théories semblent complémentaires — mais comment les articuler ? Tantôt elles se posent en concurrentes — mais comment trancher, et surtout comment, si l'on en choisit une, faire droit aux aspects que seules les autres semblent en mesure d'expliquer, de respecter, voire de seulement formuler ? Gageons que notre perplexité ne prendra pas fin de sitôt. Ayons du moins l'audace d'espérer qu'à force de suivre tantôt une voie, tantôt une autre, nous nous fassions petit à petit une idée plus riche et plus précise de ce singulier monument de la nature qu'est l'esprit juché sur le cerveau.

Daniel ANDLER,
Université Charles de Gaulle-Lille 3,
C.R.E.A., École polytechnique, Paris.

COLOR VISION : A CASE STUDY IN THE FOUNDATIONS OF COGNITIVE SCIENCE*

> *« [La couleur] est "l'endroit où notre cerveau et l'univers se rejoignent", dit [Cézanne]... Il ne s'agit donc pas de couleurs, "simulacre des couleurs de la nature", il s'agit de la dimension de couleur, celle qui crée d'elle-même à elle-même des identités, des différences, une texture, une matérialité, un quelque chose »*[1].

This statement evokes color as dimension, a place which does not pre-exist to those who gather there ; it is *brought forth* in the meeting itself. Our purpose in this paper is to address this question on the basis of our experimental work in *comparative* color vision, that is, the study of color vision in various animal species. Our intention is not to give an updated scientific review of this field, but to consider it as a case study revealing fundamental questions in cognitive science.

The paper unfolds in three stages. In the first, we briefly review some current work in the study of color vision. This view will then be taken to a critical limit in the second stage, through what we like to call the comparative argument. It purports to demonstrate the mode in which color vision is an ecologically embedded activity rather a form of information processing. We warn the reader immediately that we do *not* construe this in any way as a form of subjectivist view that color is a type of sensation, nor as a Lockean view that color is a form of secondary disposition. The comparative argument, we argue, allows to go beyond both those classical positions. This is done in the third and final part where we lay out an enactive view of color.

---

* This paper is a condensed form of a forthcoming target article for *The Behavioral and Brain Sciences*. The present paper owes much to a paper by Evan Thompson, and presented to the Dept of Philosophy, McGill University, Canada, December 1988.

1. Maurice MERLEAU-PONTY, *L'Œil et l'esprit*, Paris, Gallimard, 1964, p. 67.

I. — THE CURRENT ONTOLOGY OF COLOR

As every school child learns, modern color science owes its origins to Newton, who held that color is a sensation produced by the wavelengths that constitute light[2]. However, contrary to this received view, there is *no* one-to-one correspondence between the perceived color of a surface and the spectral composition and intensity of the light reflected from that surface. This relative independence of perceived color from spectral composition has now been extensively documented in humans.

This independence is most clearly manifested in two complementary phenomena. In the first, the perceived color remains relatively stable in spite of large changes in illumination, as is known as color *constancy*. In the second phenomenon, color *induction*, two surfaces that reflect light of the same spectral composition can be perceived to have different colors depending on the surrounding in which they are placed[3]. Given these two pervasive phenomena (and a host of concurrent evidence), it is simply futile to assimilate color to spectral composition. Such an attempt is simply evidence one does not understand the phenomena at all, since color science today is based on the very distinction between color and spectral content of light[4]. In fact to establish that an organism has color vision one tests which *combinations* of wavelengths trigger the *same* color perception.

Is there any current candidate to replace wavelength for an objectivist reduction of color? Some have proposed recently that color should be identified with surface spectral *reflectance*[5]. Reflectance is defined as the ratio at each wavelength of the incident and reflected light. Thus, it is a relatively stable, object-bound property, and it can be specified without

2. Isaac NEWTON, *Optics*, New York, Dover, 1952 (based on the 1730 ed.), p. 124-125.

3. For examples of constancy in humans, see Edwin LAND, « The Retinex Theory of Color Vision », *Scientific American*, t. 237, 1977, p. 108-128, and in a comparative context David INGLE, « The Goldfish As a Retinex Animal », *Science*, t. 225, 1985, p. 651-653; Vivian BUDNIK, Humberto MATURANA, Francisco VARELA, « Chromatic Induction : A Comparative Study », 1990, submitted.

4. For an excellent philosophical of this, see John WESTPHAL, *Colour : Some Philosophical Problems from Wittgenstein*, Oxford, Basil Blackwell, 1987, p. 75.

5. Most clearly in Robert HILBERT, *Color and Color Perception : A Study in Anthropocentric Realism*, Stanford, Center for the Study of Language and Information, 1987. Also in Patricia CHURCHLAND, « Reduction, Qualia, and the Direct Inspection of Brainstates », *Journal of Philosophy*, t. 82, 1985, p. 8-28; M. MATTHEN, « Biological Functions and Perception Content », *ibid.*, t. 85, 1988, p. 5-27.

detailed reference to the microstucture of surfaces. This in turn could allow for phenomena such as constancy and induction. The neo-objectivist argues then that color is an objective property of the world and can be identified with reflectance. Color as we perceive it, in contrast, would be indeterminate with respect to objective color since our perception does not completely specify surface reflectances ; they give us only « anthropocentrically defined kinds of colors and not colors themselves »[6].

This last phrase must be understood in the context of the current research in color vision, and more precisely in what is known as computational color vision, where the core problem is precisely to propose explicit algorithms and neural networks permitting a biological or an artificial device to *determine* surface reflectance with incomplete knowledge, that is, only with the knowledge of the spectral content of reflected light from surfaces which is all the retina ever receives[7]. We need not enter into the details of this line of work. Suffice to say it is a typical case of an ill-posed inverse problem. In essence these are essentially the three elements that enter into consideration :

1) *Low-Dimensionality* : one observes that reflectances can be described as lying within a low-dimensional manifold : only a few basic functions are enough to span the appropriate space. Correspondingly, one assumes that there are fewer parameters that describe reflectance than parameters that describe the class of photoreceptors that sample each point in the image. Thus computational color vision is fundamentally constrained by the low-dimensionality of both stimuli and receptor types. Specifically, humans have three photoreceptors, and natural reflectances need three to six basis functions[8].

2) *Global computations* : the local activity of a photoreceptor is in itself not significant. What is relevant is the global interaction over long ranges which transform luminance into « lightness », a level of activity closer to reflectance. There are a number of such equivalent « lightness algorithms »[9]. They can all be understood as a manifestation, in the living system, of the lateral interactions and reentrant circuits typical of both

6. R. HILBERT, *op. cit. supra* n. 5, p. 27.
7. For a good technical introduction, see Lawrence MALONEY, *Computational Approaches to Color Constancy*, Applied Psychology Lab. Stanford University, Technical Report 1985-01, 1985, and Anya HULBERT, *Color Computation in the Visual System*, A. I. Memo N° 814, Cambridge, MIT, 1984, and « Formal Connection between Lightness Algorithms », *Journal of the Optical Society of America A*, t. 3, 1987, p. 1684-1693.
8. L. MALONEY, *op. cit. supra* n. 7, p. 60.
9. A. HULBERT, *op. cit. supra* n. 7.

retina and visual system leading to internally specified values rather than to raw sensorial values.

3) *Segmentation* : even under low-dimension and network global computations, reflectances are still underdetermined. One key missing element is the way a scene is segmented into the relevant patches on which the calculation of reflectance will be performed. Thus, some extra assumptions about surfaces (abruptness of change, distributed averages, and so on) must be brought to bear.

The coming together of these three elements does permit to regularize the ill-posed problem of recovering reflectance with various degrees of success (or failure), and with various degrees of biological plausibility. We do not intend to go into these details.

What does this mean in terms of the current ontology for colors ? The objectivist would say that this underdetermination is precisely the basis for his views, for there are differences in reflectance we cannot detect, and hence we only have a human's eye view of objective color. This however seems to miss an important point revealed by the results of computational color vision, which amount to an intrinsic circularity in the argument. To know reflectance we must specify the dimensionality (number of receptor classes) and the surface segmentation of scene. But the objectivist assumes that color (*qua* reflectance) is an intrinsic property which helps to explain how the visual system actually segments a scene, and why evolution produced a given number of receptor classes. In other words, the three elements that go into the computation of reflectance are mutually interdependent, and do not have a hierarchy of logical precedence.

This already leaves the objectivist ontology of color in a bit of trouble. We will come back to this in the last Section. In the meantime there is more to come, and I now turn to what I call the comparative argument.

II. — THE COMPARATIVE STANDPOINT

In what we have just said about computational color vision, it is easy to lose sight of the phenomena of color in a larger context. In fact, there is a higher-level science of color which comprises elements from chemistry and physics, physiology, ecology, psychology, and artificial intelligence. We should remind ourselves of just what color *is* within these higher-level views. The existential particule stands here for the necessary components of its experiential phenomenology, to which we must give precedence.

The most common definition for color is its three-dimensional description of hue, saturation, and brightness which together define a *color space*. Consider now hue. They can be either unique or binary. A unique hue is one that is pure in the sense that it does not contain other chromatic components : blue, green... Unique hues can be opponents : blue and yellow, green and red. A binary hue, in contrast, is one that does contain other chromatic components. Some hues are necessarily binary such as orange, a mixture of red and yellow. Binary hues are always located between two unique hues in the color space.

This unique/binary structure of hues already is a symptom that something is amiss in the objectivist view of color, since we can find nothing in reflectance that has this structure. The obvious reply is to say that this unique/binary structure might be a property merely of color experience but not of objective color. But this reply places the burden back in the objectivist court, since *prima facie* color is hue, saturation and brightness as necessary properties. To say that they are only properties of experience is equivalent to claiming that colors are properties of objects, but red, blue, yellow or blue are not. Such problems show that the objectivist reduction has lost sight of the phenomena.

Let us take this direction of argumentation one step further. As we mentioned before, human color vision is trichromatic since it can be represented in a space with three independent variables. The most adequate variables are, contrary to common opinion, not the sensitivity curves of the photopigments of retinal cones, but rather some combination of them referred to as *color opponent channels*. These are well known to the physiologist and psychophysicist and are typically a non-opponent luminance (or achromatic channel), a red minus green (or tritanopic channel), and a yellow minus blue (or deuteranopic channel)[10]. These channels constitute the axes of what we will refer to as a *chromatic domain*. A chromatic domain should be distinguished from a color space, having as its dimension hue, saturation, and brightness. These dimensions specify what color is, and so specify color phenomena at their own level. A chromatic domain on the other hand has as its dimensions the functionally specified color channels of a given perceiver, which specify color at the level of its embodiment.

We can now clearly formulate the comparative argument in successive steps as follows :

---

10. For a clear presentation of the classical perspective see Leo HURVICH, *Color Vision*, Sunderland, Sinauer, 1981. For a theoretical derivation of channels, see George BUCHSBAUM, Arnold GOTTSCHALK, « Trichromacy, Opponent Colours Coding and Optimum Information Transmission in the Retina », *Proceedings of the Royal Society of London*, t. 220, 1983, p. 89-113.

(α) a chromatic domain determines a color space ;
(β) since chromatic domains are relative to the embodiment in a given perceiver class, so too is color space.

(α) It is already obvious how binary and complementary hues can be explained by appealing to the trichromacy of our color vision. We can expect a similar account for the other dimensions of color space. Indeed the brightness dimension corresponds to the axis of the achromatic channel, the hue dimension to the maximal periphery in the space spanned by the channels, and saturation to the points within the boundary of hues. This correspondence provides a bridge between the physiological properties of color vision and the phenomenological features of color space. It therefore suggests how color space is determined by our color domain.

(β) The study of color vision has been traditionally very anthropocentric. However a growing awareness is that color vision is prevalent amongst most vertebrate and many invertebrates systems, and that « the true culmination of the evolution of color vision in vertebrates is probably in the highly evolved diurnal animals perhaps best represented by diurnal birds and it is within these species that we should look for color vision significantly more complex than our own... » [11]. In fact there is strong evidence that some diurnal birds such as the pigeon and the duck are at least tetrachromats and even pentachromats [12]. There is also good evidence for tetrachromacy in fishes (such as the goldfish and the Japanese dace) [13]. The visual system of these diverse species seems to have four (perhaps five) channels, in contrast to three as in human, leading to four or five dimensional color domains [14].

Many people when they hear of this evidence respond by asking : « Well, what are the extra colors a tetrachromat or pentachromat sees ? » The question is understandable but naive. A tetrachromat for instance cannot be imagined as one who makes finer distinction say between red and yellow hues. Such an ability would be an increase in *resolution* within

11. John BOWMAKER, « Color Vision and the Role of Oil Droplets », *Trends in Neuroscience*, t. 3, 1980, p. 41-43.

12. Stuart JANE, John BOWMAKER, « Tetrachromatic Color Vision in the Duck : Microspectrophotometry of Visual Pigments and Oil Droplets », *Journal of Comparative Physiology A*, t. 162, 1988, p. 225-235 ; Adrian PALACIOS, Susana BLOCH, Carlos MARTINOYA, Francisco VARELA, « Color Mixing in the Pigeon », *Vision Research*, t. 30, 1990, p. 587-596.

13. Franz HAROSI, Yoshichi HASHIMOTO, « Ultraviolet Visual Pigment in a Vertebrate : A Tetrachromatic Cone System in the Dace », *Science*, t. 222, 1983, p. 1021-1023 ; Christa NEUMEYER, *Das Farbensehen des Goldfisches*, Hab. Thesis, Univ. Mainz, 1986.

14. For the case of goldfish and turtle, see F. VARELA, A. PALACIOS, « Tetra-chromacy : An Analysis of Color Hyperspaces », *Biological Cybernetics*, submitted.

our chromatic domain. To be a tetrachromat means that color space has an entirely new *dimension.* What could this possible mean ?

One *Gedankenexperiment* is to imagine what happens when we change unique hues starting from, say, red. In a trichromatic system, giving a circle of hues (one dimension less than the color space) a red can only move towards being yellower or being purpler, along the two axes of a line. But in a tetrachromatic system, the hue loci become a surface ; from a point identified with a pure hue, we can move in infinite directions while still remaining within the surface of pure hues. Thus there is no way to map our color experience into such a domain without a remainder. The differences between color dimensions are of the nature of *incommensurability.*

Now I have already argued that color space defines what color is. The consequence of the comparative line of analysis is then that, color as phenomena is inseparable from a variety of embodiments. Does it follow then that we should fall into a full relativity about color and treat it as a subjectivist *qualia* [15] ? Having so far tried to avoid the Scylla of objectivism, we seem now to fall into the Charybdis of subjectivism. To address this other extreme we must proceed to the second part of our comparative argument, which can be stated thus :

(γ) the structure of a given type of color perceiver implies a certain evolutionary path and ecological niche, and it is in reference to both that color is embodied.

(γ) The subjectivist position is one that abstracts itself from the mutually embedded of organisms and their milieu. To understand color fully we must understand the many variations of different types of chromatic domains. Let us illustrate with two examples. The first one is the recent observation that among new world monkeys all males are dichromats and three-quarters of the females are trichromats [16]. Several hypotheses can be invoked to account for this polymorphism : group selection (diversification of perceptual emphasis), ecological balance (diversification of sources of food). The second example is from the well-known color vision of bees, which evolved a distinctively different form of trichromacy than that of humans. Here the color domain contains opponent channels which extend into the UV range with no sensitivity in the longer wavelengths, i.e. it is displaced trichromatic domain [17].

---

15. For a recent argumentation in this sense, see C. Lawrence HARDIN, *Color for Philosophers,* Cambridge, Hacket Publishing, 1988.

16. George JACOBS, Jacob NEITZ, Michel CROGNALE, « Color Vision Polymorphism and its Photopigment Basis in a Callitrichid Monkey », *Vision Research,* t. 27, 1987, p. 2089-2100.

17. Ralph MENZEL, « Spectral Sensitivity and Color Vision in Invertebrates », in Hans AUTRUM, ed., *Handbook of Sensory Physiology,* vol. VII/6A, Berlin, Springer, 1979.

The point here is that subjectivism cannot explain these kinds of phenomena because it focuses on the individual perceiving subject, and thus ignores the appropriate ecological level of explanation. The two cases mentioned require that we see the neurobiological substrate and the relevant color channels in reciprocal specification of what counts as a niche for the organism. Notice that this argument applies with equal force to the Lockean position that color is the reflectance of objects since it refers to the naked world of physics, and not to the ecologically embedded, relevant distinction that emerge in an evolutionary history.

Thus we come to the conclusion of our comparative argument : color is always relative to the structure and history of an ecologically embodied perceiver. To explain color we must *generalize over ecologically embodied perceivers.* We now turn to discuss in more detail what we mean by the term « ecologically embodied ».

III. — AN ENACTIVE VIEW OF COLOR

It is clear that what we have called here the objectivist view of color matches well with the sensibility of cognition and perception as some form of information processing : light falling on the retina giving rise to signals which arise from a given state of affairs in the world. One weakness of this position is that, even if it puts great emphasis on internal processing, the ultimate reference point is some given environment. This completely neglects the fundamental observation that environment for a living system is not pre-given but specified and shaped along with its evolutionary history. Vision is not a re-covery of pregiven features, but a sensori-motor enactment of a possible world. A visual world is neither found nor invented, but is *enacted.* Needless to say I cannot go into a fuller explicitation of this view here, but will bring it to bear into the topic of color vision [18].

The temptation from our inertia to see perception as representation from a pre-given world is to say : « Perhaps we should conclude, as you argue, that color cannot be identified with surface reflectance, but color vision is still some more or less perfected form of recovering surface reflectance. » There are several problems with this view. First, as we

18. For more on this general line of argument, see F. VARELA, *Connaître : les sciences cognitives,* Paris, Seuil, 1988 ; H. MATURANA, F. VARELA, *The Tree of Knowledge,* Boston, New Science Library, 1987.

mentioned before, color appears in various types of niches. Deep water fishes tend to be dichromats, whereas fishes that live closer to the surface tend to be trichromats and tetrachromats[19]. What is the standard from which to judge here which is the right reflectance to recover? One could again say that this only shows that different niches represent different regularities in the external world that are optimized, as Marr would have it[20]. The weakness of this position (and a great weakness it is) is that such reliance on optimality models of the organism-environment relation lead us into a thicket of debates about the validity of optimality as an evolutionary argument[21]. We cannot pursue these problems here.

We do want however to come back to a point raised at the end of the first Section. Surface reflectances do not come ready made in the environment. How is one to specify what is an edge, a boundary and orientation, except in reference to some visual system for which these distinction are relevant? What are the relevant areas and limits that separate brilliance from glossiness? How many chromatic channels will be brought into action to specify a chromatic domain? As the recent research in A.I. shows, these are difficult, thorny issues. There is no solution except through a *finalization relative to a standard perceiving system.* This standard observer is clear for the A.I. engineer, but not for the comparative biologist : nature ranges wider. The point is not that surfaces/color/dimensions are subjective : it is that objects/color/edges simultaneously suppose a perceiver for whom something counts as one. And these come in diverse varieties.

We submit that in history the tangled webs of species have been dancing their evolutionary game in a way in which surface and color have played a central role. Such a dance has produced also various ways of coloring and there is no possibility to define a universal standard. The ways of coloring define each other like the partners in a dance. Indeed, the natural history of this poly-centered naturalized aesthetics is yet to be written.

Thus we come to the conclusion of this article, where we have attempted to show how color vision, and its comparative dimension in particular, can be seen as a case study in the foundation of cognitive science. Our conclusion is that it provides evidence for what we have called an enactive view of cognition which avoids two extremes :

---

19. Edward MACNICHOL, *The Ecology of Vision,* New York, Oxford U. Press, 1982.

20. David MARR, *Vision,* San Francisco, Freeman, 1982. See also Philip KITCHER, « Marr's Computational Theory of Vision », *Philosophy of Science* t. 55, 1988, p. 1-24.

21. For more on this point, see John DUPRÉ, ed., *The Latest on the Best,* Cambridge, MIT Press, 1987.

— First the extreme of cognition as referring to (representing) a pre-given world, even allowing for much internal processing and incompleteness.
— Second the extreme of cognition as some variety of constructivism which misses the ecological embeddedness of the organism and its constraints.

Our position is that cognition is *enaction,* a mutual en/unfoldment of organism and world revealed through regularities which are brought forth, such as color. The various ways of coloring are an excellent example : we, animals of this earth, live in our various color spaces and thus our brains and our universes meet. Cézanne would have probably loved this.

Francisco J. VARELA,
Institut des neurosciences,
C.N.R.S.-Paris VI,
C.R.E.A., École polytechnique, Paris.

Evan THOMPSON,
Department of Philosophy
University of California, Berkeley.

LE PHYSIQUE, LE MORPHOLOGIQUE, LE SYMBOLIQUE

REMARQUES SUR LA VISION

I. - CRITIQUE DU DUALISME SYMBOLIQUE/PHYSIQUE ET DU SOLIPSISME MÉTHODOLOGIQUE

1. Le dualisme symbolique/physique

1.1. Le paradigme classique — dit symbolique — des sciences cognitives actuelles est computationnel, symbolique et fonctionnaliste (pour une introduction, cf. les dossiers dans *Le Débat*, 1987 et *Préfaces*, 1988).

(i) Il postule d'abord l'existence de représentations mentales neurologiquement implémentées (et donc physiquement réalisées) dans des états mentaux. Il s'oppose sur ce point aux positions réductionnistes éliminationnistes et physicalistes qui considèrent que les représentations mentales ne sont que des artefacts de la description psychologique et ne possèdent pas d'existence objective en tant que telles (cf., par exemple, Churchland, 1984)[1].

(ii) Il postule ensuite que ces représentations mentales sont de nature symbolique, c'est-à-dire qu'elles appartiennent à un langage mental interne (le « mentalais » de Fodor) possédant la structure d'un langage formel (avec ses symboles, ses expressions, ses règles d'inférences, etc.). Il s'oppose sur ce point aux conceptions qui estiment qu'un certain nombre de résultats expérimentaux (par exemple, sur les rotations d'images mentales) plaident en faveur de représentations mentales *topologico-géométriques* non propositionnelles (cf. Kosslyn, 1980 et Shepard-Cooper, 1982).

(iii) Il postule enfin que, comme en informatique, on peut découpler les problèmes de matériel (hardware) de ceux de logiciel (software) et que les représentations mentales symboliques sont, en ce qui concerne leur structure formelle et leurs contenus informationnels, indépendantes de leur implémentation dans leur substrat physique (magnétique, neuronal,

1. Pour plus de précisions concernant les références placées entre parenthèses, dans cet article, se reporter à la Bibliographie, p. 180.

etc.). Il s'oppose, sur ce point, aux conceptions *émergentielles* qui considèrent au contraire que l'on doit concevoir ces structures formelles comme des structures stables émergeant de processus dynamiques, coopératifs et statistiques sous-jacents (cf. Thom, 1972, 1980 ; Zeeman, 1977 ; PDP, 1986 ; Smolensky, 1988 ; Petitot, 1986b, 1989f,g,i). Une épistémologie de l'émergence interroge dans le paradigme symbolique une conception formaliste et « descendante » (*top-down* en jargon) du traitement de l'information et lui oppose une conception naturaliste et « ascendante » (*bottom-up* en jargon).

Pour le paradigme symbolique, les sciences cognitives doivent par conséquent se fonder dans une théorie computationnelle des manipulations formelles de représentations symboliques. Ces représentations traitent de l'information et, en particulier, de l'information issue du monde extérieur. Elles peuvent acquérir ainsi un contenu sémantique. Mais la causalité naturelle des opérations qui agissent sur elles et leur permettent d'agir (par exemple, sur des comportements à travers des contenus intentionnels d'attitudes propositionnelles) est une causalité strictement formelle et syntaxique. Autrement dit, en tant qu'états et processus mentaux, elles sont fermées à leur sémantisme.

1.2. Le mentalisme computationnel du paradigme classique est inséparable, en ce qui concerne l'information servant d'*input* aux calculs mentaux, d'un objectivisme physicaliste standard. Selon ce dernier, ce qu'il y a d'objectif dans l'environnement se réduit à ce qu'enseigne la physique fondamentale standard : atomes, rayonnement, ondes sonores, etc. On en arrive ainsi à un véritable *dualisme* (fortement réminiscent des dualismes philosophiques traditionnels) entre *le symbolique et le physique*. Dans son ouvrage de référence *Computation and Cognition*, Zenon Pylyshyn a excellemment exposé celui-ci. L'information externe étant conçue de façon physicaliste, elle est *a priori sans signification* pour le système cognitif. Elle se trouve soumise à une transduction par des modules périphériques (ces modules comprennent les récepteurs sensoriels comme la rétine ou la cochlée mais peuvent se prolonger à des transducteurs compilés), transduction qui la convertit en information interne (fréquences de *firing* de neurones) computationnellement significative. Il existe évidemment une corrélation causale nomologiquement descriptible entre l'information physique externe et l'information computationnelle interne produite par la transduction. Mais cela n'implique pas pour autant l'existence d'une science nomologique du rapport *significatif* que le sujet entretient avec son environnement. D'une part, en effet, la transduction décrite physiquement et causalement est cognitivement opaque. Sa fonction est non symbolique. Elle fait partie de l'architecture

fonctionnelle qui contraint formellement la structure des algorithmes mentaux. D'autre part, la signification est le résultat des opérations effectuées par les représentations mentales symboliques et celles-ci ne sont pas causalement déterminées par le contenu physique objectif des états de choses externes. D'où, selon Pylyshyn, un dualisme physico-symbolique strict. Il existe une coupure irréductible entre le cognitif interne et le physique externe. Il existe un langage physique universel, cohérent et unificateur, composé de termes physiques. Mais il n'existe pas de descriptions physiques, dans ce langage, de ce qui est significatif dans l'environnement pour un sujet cognitif (cf. Petitot, 1989f). On pourrait aligner les citations concernant cette « strongest constraint » et cet « extremely serious problem » : « the relevant aspects of the environment are generally not describable in physical terms », « psychological regularities are attributable to *perceived*, not physically described properties », « the general failure of perceptual psychology to adequately describe stimuli in physical terms », etc. (Pylyshyn, 1986, p. 166-167). Il faut donc disposer de concepts perceptuels et cognitifs fonctionnels. Mais ceux-ci sont *sans* contenu physique. Le lexique physique et le lexique cognitif ne s'apparient pas naturellement. Ils ne sont compatibles qu'à travers les transductions.

On remarquera que de telles affirmations ne sont acceptables que sous certaines hypothèses :

(i) ce qui existe d'objectif dans l'environnement se réduit à ce que décrit la physique fondamentale standard ;

(ii) ce qui est significatif doit, pour être significatif, être au préalable représenté ;

(iii) la représentation s'identifie à un calcul : l'esprit est computationnel.

1.3. Comme l'ont noté de nombreux auteurs (Putnam, Searle, Dreyfus, etc.), deux grands problèmes demeurent énigmatiques dans le paradigme classique.

(i) Du côté du sujet, le problème *du sens et de l'intentionalité.* Comment des représentations mentales symboliques peuvent-elles acquérir un sens, une interprétation, une dénotation, une orientation intentionnelle vers le monde externe ? Comment un système cognitif peut-il agir en fonction du sens des symboles et des expressions symboliques alors qu'il ne possède de relations causales qu'avec la forme (logico-syntaxique) de ceux-ci ? Il ne suffit pas de dire que le sens est le résultat d'une « interaction » sujet-monde puisque cette interaction n'est pas nomologiquement descriptible et explicable.

(ii) Du côté du monde, le problème *de la manifestation qualitative et*

morphologique des phénomènes. Comme Ray Jackendoff y a beaucoup insisté, on ne peut se borner à poser que le monde phénoménologique de l'expérience est un simple résultat des opérations de « l'esprit computationnel » (Jackendoff, 1987). Encore faut-il comprendre la part de ces opérations, en général opaques pour la conscience phénoménologique, qui se trouve devenir constitutive de la structuration qualitative du monde en choses, états de choses, événements, processus, etc., perceptivement appréhendables et linguistiquement descriptibles. En effet, le processus computationnel est inconscient. Seules quelques-unes des structures qu'il produit sont conscientes. On peut alors, comme Jackendoff, adopter un point de vue « projectiviste » faisant du monde phénoménologique un monde « projeté » résultant d'une « projection » de constructions cognitives, poser que la plus grande partie de la structure interne des constituants du langage mental (ce que Jackendoff appelle la « structure conceptuelle ») n'est pas projetable et faire de la « conscience » phénoménologique (différente, donc, de l'esprit computationnel) le corrélat (en un sens proche de celui de la corrélation noèse/noème chez Husserl) de ce monde projeté (le « Mind-Mind problem »). Mais on peut également, comme nous le ferons plus bas, utiliser les résultats scientifiques théoriques et expérimentaux qui démontrent l'existence de structures morphologiques et qualitatives *objectives émergeant*, par un processus dynamique (auto)organisateur, des substrats physiques. Ce point de vue proprement *« morpho-génétique »* s'oppose au point de vue *« morpho-projectif »*. Il prend appui sur l'existence démontrée d'un niveau de réalité morphodynamique que l'on pourrait appeler un niveau *« phéno-physique »* (expression phénoménologique du niveau de réalité proprement physique à travers un processus *naturel* objectif, non cognitif, de phénoménalisation des substrats matériels).

2. Les limites épistémologiques du cognitivisme symbolique : la non-prise en compte de la dimension morphodynamique

2.1. Fortement tributaire des recherches en Intelligence Artificielle (IA) dont il a hérité du point de vue computo-représentationnel, le cognitivisme symbolique, dans son rapport aux neurosciences, à la psychologie et à la philosophie de l'esprit, a permis des progrès décisifs dans la compréhension et la formalisation des mécanismes mentaux constitutifs du « sens commun » (applications de règles en fonction du contexte, inférences, décisions, représentations des connaissances, rôle causal du contenu intentionnel des attitudes propositionnelles dans le comportement et l'action, etc.). Se voulant science des états et des processus mentaux, son projet est de comprendre les sujets cognitifs en tant que

« systèmes symboliques physiques » et de *naturaliser* l'esprit, le langage et le sens.

Pour comprendre à quel point son statut épistémologique est toutefois délicat et problématique, il suffit de remarquer qu'il reprend l'ensemble des problèmes de la tradition sémantique (logique, philosophie analytique, etc.) en termes computationnels, qu'il les relie aux neurosciences et que, sur cette base, il transforme les descriptions noético-noématiques de l'expérience phénoménologique en sciences naturelles.

On peut s'étonner par conséquent du fait que, dans l'ensemble des débats (fort vifs) qui se sont développés à son sujet, les concepts ontologiques, théoriques et épistémologiques les plus fondamentaux — comme ceux de matière, de réalité physique, d'idéalité symbolique, de causalité, etc. — soient utilisés de façon non critique dans leur acception souvent la plus banale.

Par exemple, une des raisons principales du rejet des conceptions émergentielles par le cognitivisme symbolique vient d'une incompréhension de l'épistémologie de l'émergence. Lorsqu'un système est un système à *deux* niveaux d'organisation, par exemple un niveau qualitatif « macro » et un niveau physique « micro » sous-jacent, le niveau supérieur « macro » est causalement (au sens de la causalité matérielle) réductible au niveau inférieur. Mais cela ne l'empêche évidemment pas de posséder des éléments de structure très largement indépendants de la structure fine « micro » sous-jacente. Ces éléments possèdent une certaine autonomie objective. Cela est tout à fait banal en physique (phases, transitions de phases, défauts dans les cristaux liquides, etc.). Comme l'a souligné Searle, ce n'est que si l'on identifie un phénomène à sa genèse causale — autrement dit, si l'on passe subrepticement d'un réductionnisme causal, justifié, à un réductionnisme ontologique matérialiste, dogmatique et donc injustifié — que l'on est conduit à dénier l'autonomie et la réalité objective des niveaux supérieurs.

De même, lorsque certains auteurs s'essaient à dépasser le dualisme du physique et du symbolique pour développer un monisme naturaliste, ils le font en général à partir d'un matérialisme vulgaire ou d'un physicalisme ne tenant aucun compte de récents résultats fondamentaux de certaines disciplines physiques. Par exemple, on cherchera à développer un behaviorisme physicaliste faisant des contenus mentaux de simples réponses de l'organisme à des états de choses. Ou bien on posera, au contraire, l'identité entre les états mentaux et des états cérébraux, quitte à affronter les multiples difficultés qui en découlent.

De même encore, pour en revenir au dualisme, le *solipsisme méthodologique* est la conséquence directe d'une certaine conception de l'objectivité physique. Selon Fodor par exemple, il est impossible d'introduire

dans une psychologie scientifique le rapport significatif qu'un sujet cognitif entretient avec son environnement. En effet, ce rapport n'est pas, nous l'avons vu, nomologiquement légalisable dans l'état actuel des connaissances. On ne pourrait donc l'introduire qu'en termes, non scientifiques, de sens commun. D'où la légitimité de la morale provisoire solipsiste : seuls les contenus « étroits » (*de dicto* et non *de re*, ne dépendant que du sujet, de son langage mental interne et non pas de sa relation contextuelle à l'environnement) interviennent dans l'individuation et l'identification des états mentaux et possèdent des capacités causales (cf. Jackendoff, 1987).

2.2. Tout cela pour dire que l'ensemble du débat actuel sur la cognition dépend de façon déterminante de la préconception que se font les cognitivistes de l'objectivité physique. Un de leurs préjugés fondamentaux est *qu'il n'existe pas de physique qualitative des formes, de physique morphologique, de phéno-physique.* Or ce préjugé n'est justifié que pour la physique fondamentale (relativité générale et mécanique quantique incluses). *Il ne l'est absolument plus* si l'on prend en compte les résultats, profonds, nombreux et convergents, de l'ensemble des disciplines physiques qui se sont intéressées ces dernières années aux phénomènes d'(auto)organisation des substrats matériels.

Nous avons longuement commenté ailleurs ces travaux mathématiques et physico-mathématiques remarquables (cf., par exemple, Petitot, 1982, 1986b, 1989g et, surtout, leurs bibliographies) : théorie qualitative de la structure et de la stabilité structurelle des systèmes dynamiques non linéaires, de leurs attracteurs et de leurs bifurcations, attracteurs étranges et chaos déterministe, théorie des singularités et de leurs déploiements universels, théorie des phénomènes critiques (transitions de phases, etc.) et des phénomènes de rupture de symétrie dans les phases mésomorphes, structures dissipatives, etc. Ces résultats ont montré expérimentalement et démontré mathématiquement que, dans de nombreux systèmes naturels organisés à (au moins) deux niveaux (cf. plus haut), le niveau « macro » (global, grossier, en général finiment descriptible) émergeant, à travers des comportements collectifs ordonnés et coopératifs, du niveau « micro » sous-jacent (local, complexe, en général non finiment descriptible) est essentiellement organisé autour des *singularités* des processus physiques « micro ». Les singularités *structurent morphologiquement* les phénomènes. Elles sont *phénoménologiquement dominantes* et soumises à des contraintes *abstraites et universelles* (« platoniciennes ») mathématiquement formulables et largement indépendantes de la physique « fine » des substrats.

Le concept de physique qualitative des formes, de physique morphologique, de phéno-physique, *appartient désormais au concept de réalité*

objective. Ce fait a, selon nous, des conséquences incalculables, à la fois théoriques et épistémologiques, pour le cognitivisme. En effet, comme nous le verrons, *le morphologique constitue un moyen terme entre le physique et le symbolique* : il est d'origine physique (émergent) mais sans être pour autant matériel, il est formel mais sans être pour autant symbolique ; il est *topologiquement et géométriquement* formel et non pas *logiquement* formel. Sa prise en considération rend légitime la double hypothèse suivante :

(i) il existe une information morphologique et qualitative présente dans le monde externe qui, tout en étant d'origine physique, est néanmoins de nature phénoménologique et, à ce titre, intrinsèquement significative ;

(ii) cette information morphologique est reconstituée après transduction et sert de base aux processus proprement symboliques de traitement de l'information.

Selon nous, la plupart des difficultés (voire des apories et des paralogismes) du cognitivisme classique proviennent du fait qu'il cherche à engendrer le morphologique à partir d'une conception logico-combinatoire (somme toute encore logiciste et analytique) du syntaxique et du sémantique alors que cela est pourtant clairement impossible, puisque les dimensions intrinsèquement spatio-temporelles et dynamiques du morphologique *ne sont pas* d'ordre formel au sens logico-symbolique (bien que physiquement réalisées, elles ne sont pas non plus d'origine physique). Comme y insiste Jackendoff, des représentations sémantiques propositionnelles ne peuvent pas être mises au fondement d'une expérience des formes.

2.3. Le problème philosophique qui intervient ici est considérable (cf. Petitot, 1982, 1986a, 1989f). Notre propos n'est pas de le reprendre. Mais nous ne saurions trop insister sur la limite fondamentale que constitue l'orientation dogmatiquement propositionaliste du cognitivisme symbolique. Une telle orientation n'est, en effet, légitime que dans le cadre d'un objectivisme logique, d'une sémantique formelle et/ou d'une logique phénoménologique des essences. Elle est *incompatible* avec une thèse *naturaliste* quelle qu'elle soit, car il n'existe pas de formes symboliques dans la nature externe ou interne. Il ne peut exister tout au plus que des formes géométriques et dynamiques. Toute naturalisation de l'esprit, du langage et du sens présuppose donc une révolution dans la conception du formel héritée du formalisme logique. Elle présuppose catégoriquement que les formes de l'esprit, du langage et du sens soient des formes géométriques et dynamiques. Ces formes doivent évidemment être symboliquement traductibles et manipulables à des niveaux cognitifs supérieurs de représentation. Mais leur *type d'objectivité* ne peut pas, pour des raisons de principe, être originairement celui de l'objectivité symbolique.

Disons brièvement que, si elle est *naturelle*, la « formellité » de l'esprit, du langage et du sens ne peut pas être symbolique. Pour la décrire et l'expliquer, il faut passer en quelque sorte d'une symbologie à une topologie.

Paraphrasant un aphorisme de Kant (« les intuitions sans concepts sont aveugles et les concepts sans intuitions sont vides »), on pourrait dire que le cognitivisme symbolique est « aveugle » et « vide » dans la mesure où il n'arrive pas à élaborer une authentique phénoménologie de la perception. En vérité, aucun passage du physique au symbolique n'est envisageable tant que l'on ne tient pas compte du fait :

(i) que le physique est spatio-temporellement conditionné (ce que Kant appelait l'Esthétique transcendantale) ;

(ii) que ce conditionnement spatio-temporel de la physique fondamentale est prolongeable aux dimensions topologiques, géométriques et dynamiques de la phéno-physique morphologique ;

(iii) que le symbolique constitue un niveau formel de surface par rapport aux infrastructures morphologiques.

3. La thèse de la morphodynamique cognitive et le principe de double émergence

Les thèses sous-jacentes à notre réflexion sont donc les suivantes.

(i) Entre le physique et le symbolique il existe la médiation du morphologique. Sans elle, il est impossible de dépasser le dualisme du physique et du symbolique et d'accéder à une théorie naturaliste intégrée (moniste mais non réductionniste) de leur unité ontologique.

(ii) Les structures morphologiques sont de façon générale les produits de processus dynamiques d'organisation des susbtrats (physiques ou mentaux). Elles émergent des substrats et sont phénoménologiquement dominées par les discontinuités qualitatives issues des singularités, des bifurcations, des instabilités structurelles, de ces processus dynamiques.

(iii) Les structures qualitatives émergentes existent aussi bien du côté du sujet cognitif que du côté du monde naturel.

(iv) L'information morphologique résiste à la transduction. Elle est encodée dans, et véhiculée par, les signaux lumineux et sonores, puis décodée-recodée par les transducteurs. Mais, au cours de cette opération, elle se reconstitue en restant en grande partie isomorphe à elle-même. Les discontinuités qualitatives sont « contagieuses » : elles se transfèrent de substrat à substrat.

Du côté du sujet cognitif, le programme de recherche d'une morphodynamique a pour vocation de développer une idée maîtresse introduite par R. Thom et Ch. Zeeman il y a déjà plus d'une vingtaine d'années, à savoir

qu'une unité sémantique est identifiable à *la topologie d'un attracteur* d'une dynamique neuronale sous-jacente et que les structures combinatoires et logico-algébriques des automatismes de la compétence doivent par conséquent être interprétées comme des régularités émergentes stables. Cette idée a été extensivement développée en sémio-linguistique par l'école morphodynamique (cf. Thom, 1972, 1980, 1988 ; Wildgen, 1982 ; Brandt, 1986 ; Petitot, 1977, 1979, 1982, 1983, 1985, 1988, 1989a, c, d, f). Elle a été également — et indépendamment — développée dans les modèles connexionnistes du paradigme dit sub-symbolique (cf., par exemple, PDP, 1986 ; Smolensky, 1988 ; Amit, 1989). Le principal apport de ces modèles plus récents est d'avoir explicité les dynamiques « concrètes » qui intervenaient dans les modèles morphodynamiques généraux. Cela permet de spécifier ce que l'on entend par « substrat mental ». Mais, à part cela, les principaux concepts dynamiques du connexionnisme (attracteurs, bassins d'attraction, fonctions de Liapounov, stabilité structurelle, bifurcations d'attracteurs, quasi-attracteurs, ruptures de symétrie, dynamiques rapides et dynamiques lentes, phénomènes coopératifs et propriétés émergentes, etc.) sont les concepts de dynamique qualitative, de théorie de la bifurcation, de théorie des singularités, de thermodynamique statistique et de théorie des phénomènes critiques que les modèles morphodynamiques avaient déjà transférés (d'ailleurs dans l'incompréhension la plus générale) dans le domaine des disciplines psychologiques et sémio-linguistiques au début des années 1970.

Du côté du monde naturel, le programme de recherche d'une morphodynamique a pour vocation d'étudier les processus de phénoménalisation des substrats matériels (externes, non internes), de théoriser mathématiquement l'information morphologique qui en émerge, de comprendre comment cette information morphologique se trouve encodée dans, et véhiculée par, les signaux lumineux et sonores.

Ayant traité ailleurs des relations entre la morphodynamique et le connexionnisme (Petitot, 1989f, i), nous nous focaliserons ici sur le problème *du type mathématique de l'information morphologique*. La possibilité d'élaborer une *phénoménologie de la perception* satisfaisante constituant un enjeu décisif dans les débats que nous avons évoqués, nous nous limiterons à l'exemple de la perception visuelle. De façon à pouvoir être suffisamment précis tout en demeurant à l'intérieur de limites raisonnables, nous nous bornerons à *un* problème *très particulier* (mais fondamental), celui de la reconstruction des objets à partir de leurs contours apparents. Qui plus est, nous dialoguerons avec des théories particulières, mais généralement acceptées (bien que parfois controversées sur certains points), nommément celles de David Marr et de Jan Koenderink. Cela nous permettra d'expliciter certaines des thèses proposées.

II. - INFORMATION MORPHOLOGIQUE ET THÉORIE DES SINGULARITÉS EN PERCEPTION VISUELLE

Des quatre domaines fondamentaux des sciences cognitives : perception, langage, inférence, action, nous choisissons donc, pour notre exemple, le premier. Des deux points de vue traditionnels : celui concernant le développement et celui concernant l'organisme adulte, nous choisissons le second. Des quatre niveaux d'analyse : biologique (mécanismes neurophysiologiques), psychologique (processus fonctionnels de détection, représentation, stockage, utilisation finalisée d'informations, etc.), computationnel (modélisation algorithmique), mathématique (propriétés formelles de la compétence), nous choisissons le troisième et le quatrième, mais dans une optique non symbolique. Nous allons, en fait, esquisser de façon brève et relativement peu technique quelques éléments de morphodynamique qui permettent d'analyser mathématiquement les contraintes topologiques, géométriques et optiques qui contraignent de façon essentielle la formation des images visuelles et leur traitement computationnel.

1. Processus modulaires et processus centraux. Traitement ascendant et traitement descendant

La vision computationnelle est la discipline théorique qui élabore des modèles mathématiques pour les processus de construction de représentations tridimensionnelles (3D) distales à partir d'images rétiniennes bidimensionnelles (2D) proximales. Elle doit donc élucider théoriquement et modéliser mathématiquement :

(i) les processus physiques de constitution de scènes externes morphologiquement organisées ;

(ii) les processus optiques d'encodage et de propagation de ces informations morphologiques ;

(iii) le processus physico-géométrique de formation des images par projection ;

(iv) le processus sensoriel périphérique d'analyse du signal (transduction) ;

(v) la façon dont l'information morphologique ainsi décodée et recodée contraint de façon essentielle la construction des représentations ;

(vi) les rapports (par exemple de compilation) entre les niveaux suc-

cessifs de représentation (du topologico-géométrique vers le symbolique) ;

(vii) la façon dont les représentations de niveau supérieur (3D et au-delà) possèdent ou non un contenu *objectif*.

Il existe au moins deux grandes conceptions de la vision computationnelle. Pour les expliciter brièvement, reprenons l'opposition fodorienne entre processus périphériques modulaires et processus centraux non modulaires (cf. Fodor, 1984). La thèse est qu'il existe (au moins) deux types très différents de systèmes cognitifs. Les premiers sont les systèmes périphériques modulaires. Ils ont pour fonction de transformer les informations neuronales périphériques fournies par les transducteurs en représentations possédant un format propositionnel adéquat pour les calculs symboliques mentaux. Ce sont des transducteurs compilés, fonctionnant automatiquement et de façon strictement ascendante (« bottom-up » : du périphérique vers le central) comme des réflexes computationnels. Ils sont spécifiques et informationnellement cloisonnés (c'est-à-dire insensibles aux croyances, aux connaissances, aux attentes, etc., du sujet). Ils formulent des hypothèses et effectuent des inférences non démonstratives permettant aux stimuli sensoriels proximaux d'être transformés en représentations sur des objets distaux.

Mais il y a également les systèmes cognitifs centraux, qui sont non modulaires, non spécifiques, non cloisonnés, descendants, interprétatifs (et donc sensibles aux croyances, connaissances, attentes, etc.). Dans la mesure où il n'existe aucun contrôle nomologique de leur fonctionnement, ils ne sont pas, selon Fodor, traitables scientifiquement : c'est le problème du holisme sémantique. Ils sont « isotropes » (toute croyance, toute connaissance, toute attente est partiellement pertinente pour le traitement et l'interprétation de toute sortie des modules) et « quiniens » (l'ensemble des croyances, etc., influe sur chaque traitement, etc.). D'où d'ailleurs, chez Fodor, une critique de l'Intelligence Artificielle et des systèmes experts qui traitent les systèmes centraux *comme si* ils étaient modulaires, spécifiques, non isotropes et non quiniens.

Un des aspects du holisme sémantique est précisément le solipsisme méthodologique débattu plus haut.

Dans une approche « descendante » (« top-down ») inspirée de l'IA, on considère que le traitement de l'information rétinienne se réduit essentiellement à des processus *d'interprétation* des images, processus *inférentiels* régis par des connaissances. Mais une telle approche n'est pas directement applicable à la vision *naturelle*. Pour celle-ci, l'environnement est trop complexe, trop fluctuant et trop peu contraint pour être traitable à partir de mécanismes de détection de traits et d'applications de règles. Dans la vision naturelle, il existe une partie considérable du

traitement de l'information qui est modulaire et « ascendante » (« bottom-up »). Plusieurs modules fonctionnels spécifiques, indépendants et fonctionnant en parallèle coopèrent dans le processing visuel précoce et leur produit intégré sert de base aux niveaux supérieurs (centraux) de représentation et d'interprétation.

La théorie de David Marr qui nous servira de base de discussion est modulaire et ascendante. Comme l'explique le collègue de Marr, Tomaso Poggio, elle considère que la tâche centrale de la vision computationnelle est de résoudre un *problème inverse*. Il existe un processus de projection des scènes 3D sur des images 2D. Le problème inverse est celui de la reconstruction des scènes 3D à partir des images 2D. *Mais l'on voit que ce problème inverse est double, à la fois cognitif et objectif.* Il est *objectif* dans la mesure où l'on peut le traiter de façon purement géométrique et optique, sans aucune référence à un esprit computationnel. Il est également *cognitif* dans la mesure où l'on peut le traiter en termes computationnels. La thèse est que *le problème inverse objectif contraint et finalise le problème inverse cognitif.* Autrement dit, il est impossible d'expliciter les algorithmes de la vision computationnelle si l'on ne connaît pas au préalable de façon précise le *type mathématique* des structures informationnelles à traiter.

Un tel point de vue est *néo-écologique*. Rappelons que l'on appelle « écologisme » la thèse *réaliste* de James Gibson selon laquelle, dit en termes plus actuels :

(i) il existe dans l'environnement des structures qualitatives et cognitivement significatives qui sont objectives sans être pour autant strictement physiques (ce que nous avons appelé le phéno-physique) ;

(ii) le système visuel détecte et extrait ces invariants phéno-physiques et construit sur cette base objective ses inférences et ses interprétations.

L'écologisme s'oppose au solipsisme méthodologique. Selon lui, les représentations symboliques représentant l'information ont pour fonction *d'expliciter* certains aspects de celle-ci.

2. Les trois niveaux de la théorie de Marr et leurs corrélats objectifs

La théorie de Marr concerne la vision computationnelle. On en trouvera une analyse conceptuelle et épistémologique dans Kitcher, 1988. Pour une introduction générale à la théorie de la vision, on pourra consulter, par exemple, les excellents Pinker, 1984, Brady, 1982, Ballard-Brown, 1982, Ullman, 1984, Stillings *et al.*, 1987.

Selon Marr, la « quintessence » de la vision comme traitement d'information est d'extraire, par corrélation, de l'information sur les objets du

monde objectif externe à partir de la façon dont la lumière réfléchie par les surfaces physiques engendre des patterns 2D I(x,y) de luminance. A travers la transduction rétinienne effectuée par les photorécepteurs, ces patterns se trouvent discrétisés (comme les pixels d'un écran). La seule information *explicite* est, à l'entrée du système, I(x,y). A la sortie opèrent des représentations de haut niveau effectuant des tâches cognitives supérieures : différenciation d'objets, repérage de positions, appréhension de mouvements, perception des dimensions, formes et textures des surfaces, reconnaissance d'objets, regroupement par classes de ressemblance (catégorisation), etc. Comment s'opère donc, dans une théorie ascendante comme celle de Marr, le passage vers ce que G. Miller appelait « the crowning intellectual accomplishment of the brain », à savoir le monde réel ?

Marr introduit plusieurs niveaux de représentation explicitant certains aspects de l'information encodée dans les patterns I(x,y). Parmi ceux-ci trois sont fondamentaux.

(i) Le premier niveau, dit niveau 2D du « primal sketch » ou de *l'esquisse primaire,* est celui du traitement du signal I(x,y). Il s'agit *d'en expliciter la morphologie et l'organisation géométrique* de façon à pouvoir opérer des segmentations qui serviront de support aux phases intermédiaires et aux phases finales, cognitives et inférentielles, d'interprétation, de reconnaissance, de compréhension, etc. Ce niveau se décompose lui-même en (au moins) deux sous-niveaux.

(i)-a. Au niveau du « raw primal sketch », il s'agit essentiellement d'une analyse *locale* du pattern I(x,y) en termes de discontinuités qualitatives : segments de bords, terminaisons de bords, discontinuités d'orientation de bords (coins), petits domaines fermés (« blobs »), petits segments de barres, etc.

(i)-b. Au niveau du « full primal sketch », ces éléments locaux (souvent en mouvement) se trouvent agrégés et organisés, ce qui engendre des effets gestalistes bien connus : bords virtuels, etc.

(ii) Le second niveau, dit niveau 2-1/2D (pour bien montrer qu'il est intermédiaire entre le niveau 2D et le niveau 3D), est le niveau essentiel de la théorie de Marr. Nous y reviendrons plus loin. C'est un niveau unitaire globalement organisé où convergent et s'intègrent plusieurs computations modulaires effectuées sur l'esquisse primaire : les contours des surfaces visibles, les textures, la stéréopsie, le mouvement, l'ombrage, etc. Il représente le monde externe comme composé de surfaces visibles remplies de qualités sensibles et se déplaçant dans $\mathbb{R}^3$. Il n'est ni sensoriel (puisque les surfaces sont distales) ni objectif (puisque les apparences sont encore subjectives). C'est le niveau de *l'apparaître phénoménologique*. Comme nous allons le voir, il est d'essence proprement *morphologique*.

(iii) Le troisième niveau, dit niveau des modèles 3D, est celui, proprement objectif, des choses réelles, des volumes matériels et de leurs propriétés réales. C'est à partir de lui qu'opèrent les tâches cognitives supérieures et les constituants de la structure conceptuelle au sens de Jackendoff, par exemple la décomposition hiérarchique de formes en parties, la constitution de prototypes, etc. On peut faire l'hypothèse que la perception est un processing ascendant « 2D→2-1/2D→3D→ Structure conceptuelle » possédant des feed-back descendants (anticipations, inférences, interprétations, etc.) « Structure conceptuelle →3D→2-1/2D ». Le niveau 2-1/2D serait donc la fin du processing perceptif proprement ascendant. Comme le dit Marr, *c'est celui de la « perception pure »* (d'où son importance).

On remarquera que les niveaux 2D et 3D possèdent des corrélats objectifs. Les corrélats objectifs (non cognitifs) du niveau 2D relèvent, par exemple, de l'optique ondulatoire, de la photométrie, de l'analyse spectrale et de l'analyse de Fourier, de la théorie du signal, etc., c'est-à-dire des théories physico-mathématiques permettant de comprendre la formation d'images. Les corrélats objectifs (non cognitifs) du niveau 3D sont non moins évidents. Ils relèvent par exemple de la géométrie de l'espace, de la structure du groupe de Lie SO(3) des rotations de $\mathbb{R}^3$, de la mécanique du mouvement des solides, de la représentation des volumes, etc. Et il est clair que les théories objectives de ces corrélats objectifs contraignent et finalisent les algorithmes opérant sur ces niveaux puisqu'elles *déterminent* le *type* de l'information qui doit être explicitée et la nature des tâches computationnelles à effectuer.

Or, curieusement, on n'admet pas en général que le niveau 2-1/2D puisse posséder également des corrélats objectifs. Toujours fidèle au dualisme physique/symbolique, on postule une simple complémentarité entre le traitement numérique de l'image (analyse du signal et théorie de l'information) et son interprétation symbolique (structures sémantiques, inférences, etc.). Entre le numérique et le sémantique, on n'introduit pas en général ce qui est pourtant le caractère le plus manifeste de la perception visuelle, à savoir d'être une perception de formes. Cela est d'autant plus étrange que les théories géométriques qui permettent d'analyser les formes comptent parmi les plus profondes, les plus vastes et les plus prestigieuses de toute la géométrie. *Cette méconnaissance théorique constitue selon nous la limite principale des théories actuelles de la vision computationnelle.* Notre thèse est que :

(i) le niveau 2-1/2D de Marr *possède bien pour corrélat objectif un niveau de réalité* ;

(ii) ce niveau est précisément le niveau *morphologique* de la « phénophysique » ;

(iii) la théorie *objective* (physico-mathématique) de ce niveau —

théorie qui existe — contraint donc et finalise de façon essentielle tous les algorithmes envisageables au niveau 2-1/2D.

3. Le niveau 2D et le concept de discontinuité qualitative

La façon dont Marr conçoit le niveau 2D de l'esquisse primaire est exemplaire de sa conception. A ce niveau se nouent trois dimensions :

(i) les données de la neurophysiologie ;

(ii) le traitement du signal (transduction) ;

(iii) la finalisation des algorithmes rétiniens par le problème inverse objectif (au sens exposé plus haut).

3.1. Les données de la neurophysiologie.

Rappelons très brièvement et très sommairement quelques éléments de la structure générale du système visuel (cf. Buser et Imbert, 1987).

La rétine réalise une énorme *compression* de l'information visuelle et cela essentiellement grâce à l'organisation *antagoniste* centre-périphérie des champs récepteurs des *cellules ganglionnaires* dont les axones constituent le nerf optique. Ces neurones visuels sont sous-jacents aux photorécepteurs superficiels. Ils répondent essentiellement aux discontinuités. La compression de l'information rétinienne fait passer d'environ 160 millions de photorécepteurs à environ 1 million de fibres dans le nerf optique. L'image est ainsi traitée de façon modulaire et organisée en traits distinctifs (arêtes rectilignes contrastées, courbure de bord, mouvement d'un contour selon une direction donnée). Il existe des champs de fibres — des modules — spécialisées dans certaines opérations et opérant sur l'ensemble de la rétine. D'où une *cartographie* du message rétinien relayée avec une bonne *rétinotopie* (une bonne préservation des relations topographiques) jusqu'au cortex visuel primaire. Le relais fondamental est *le corps genouillé latéral* dont les cellules sont analogues aux ganglionnaires rétiniennes et encore plus sensibles au contraste local. Les représentations cartographiques s'y superposent en couches (en registres). D'où une organisation modulaire en *colonnes*, dites *colonnes de projection*, associées à une même zone du champ visuel. Les différences d'organisation et de physiologie des cellules rétiniennes se traduisent dans ces structures supérieures post-rétiniennes par l'innervation de couches différentes. Les opérations des différentes classes fonctionnelles de cellules rétiniennes (en particulier ganglionnaires) sont donc maintenues *séparées* (modularité).

Après le corps genouillé latéral, les radiations optiques traversent la substance blanche et arrivent à l'aire visuelle primaire occipitale (aire

striée) : aire principale 17 et aires secondaires 18 et 19. Les colonnes genouillées sont projetées avec préservation de la rétinotopie. Le cortex strié est organisé lui aussi modulairement en colonnes (superposition de couches, cf. les travaux de Hubel et Wiesel) ce qui permet de représenter avec une bonne rétinotopie sur la *surface* du cortex non seulement la *position* dans le champ visuel mais également *d'autres* variables comme la dominance oculaire et l'orientation. L'existence de colonnes de dominance oculaire et de colonnes d'orientation dont les ensembles sont *indépendants* et *transversaux* l'un à l'autre implique que l'aire primaire soit décomposée en *hypercolonnes* (d'environ 1 mm^2 de section) dont chacune traite les contours contrastés (les discontinuités qualitatives) dans toutes les directions de vision binoculaire d'un domaine spatial. On peut donc faire l'hypothèse que le cortex strié visuel sert à extraire de façon *topographique* des attributs visuels caractéristiques et stables comme la couleur, l'orientation, la direction, la vitesse. Ces attributs seraient alors *redistribués* de façon globale (*non* topographique) dans les aires secondaires afin d'y être analysés.

La transduction s'opère au niveau des photorécepteurs, évidemment au moyen d'intermédiaires photochimiques. Des pigments rétiniens (chromoprotéines comme la rhodopsine) absorbent l'énergie lumineuse dans les récepteurs photiques. Leur isomérisation déclenche une chaîne d'événements dans le cytoplasme de ces récepteurs, chaîne aboutissant au blocage du courant dans la membrane plasmique et, donc, à une variation du potentiel membranaire.

La rétine contient, entre les photorécepteurs et les cellules ganglionnaires, d'autres couches de cellules (bipolaires, horizontales, amacrines). En ce qui concerne l'analyse morphologique des stimuli (la couleur pose d'autres problèmes), c'est *l'organisation spatiale des champs récepteurs* (c'est-à-dire la surface de l'espace visuel et de la rétine à laquelle une cellule réagit) *qui est essentielle.* La plupart des neurones rétiniens possèdent une organisation *concentrique et antagoniste* de leur champ récepteur. Ils sont, par exemple, Centre-ON et Périphérie-OFF si un stimulus lumineux ponctuel appliqué au centre du champ récepteur conduit à une activation du centre et à une inhibition de la périphérie.

Les cellules ganglionnaires sont essentielles car elles constituent le terme de la transduction. C'est à travers elles (à travers leurs axones constituant, nous l'avons vu, le nerf optique) qu'est transmis le message rétinien aux niveaux post-rétiniens. Elles sont ON, OFF ou ON-OFF et, en ce qui concerne leur réponse temporelle, soit *toniques* (répondant pendant toute la durée du stimulus), soit *phasiques* (répondant seulement à une discontinuité temporelle du stimulus). Elles se regroupent en trois classes fonctionnelles principales X, Y et W. Les cellules X sont énergétiques et toniques. Leur gradient d'antagonisme centre/périphérie est fort,

leur résolution spatiale élevée et leur résolution temporelle faible. Ce sont des analyseurs de contrastes spatiaux, et donc de formes. A l'inverse, les cellules Y sont des détecteurs de mouvements et des analyseurs de structures temporelles.

3.2. L'Analyse du signal : critère de zero-crossing et ondelettes.

Évidemment, il existe des interactions subtiles et compliquées entre les différents neurones rétiniens : mécanismes de renforcement et d'inhibition latérale, combinaisons de contrastes spatiaux et chromatiques, etc. Mais l'on voit néanmoins apparaître clairement un certain nombre de faits massifs. Le plus massif est sans doute que, de par la structure de leur champ récepteur et leur caractère tonique, les cellules X ont pour fonction de détecter des contrastes, c'est-à-dire des discontinuités qualitatives de la luminance.

Marr a formalisé ce contenu fonctionnel en introduisant le dispositif de détection de discontinuités qu'il a appelé le critère de *zero-crossing*. L'idée en est simple.

Considérons une fonction différentiable d'une variable réelle f(x). La traversée d'une discontinuité se caractérise par un pic de la dérivée première (distribution δ de Dirac) et par un double pic — un pic positif et un pic négatif séparés par une traversée de 0, c'est-à-dire un « zero-crossing » — de la dérivée seconde (cf. figure 1).

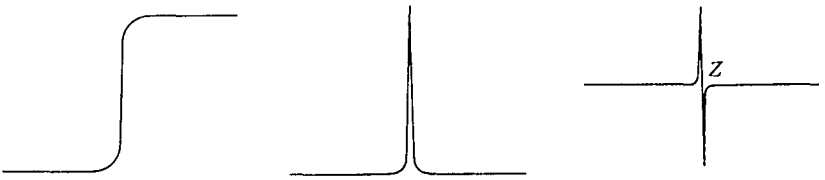

Figure 1. *Le critère de zero-crossing (d'après Marr, 1982, p. 54). (a) discontinuité de la fonction f, (b) pic de la dérivée f', (c) double pic de la dérivée seconde f''.*

Il s'agit de généraliser à deux dimensions. Pour ce faire on va :

(i) lisser localement le pattern d'intensité I(x,y) à une certaine échelle, par exemple en opérant une convolution $G*I$ avec une gaussienne centrée en un certain point : $G(r) = \exp(-r^2/2\pi\sigma^2)$ (r = distance au point considéré) ;

(ii) considérer les dérivées secondes, c'est-à-dire le laplacien $\Delta(G*I)$.

Marr remarque alors les deux choses suivantes :

(i) Comme $\Delta(G*I) = \Delta G*I$, on peut effectuer la double opération de lissage et de dérivation *en effectuant la convolution du signal avec le laplacien d'une gaussienne.*

(ii) Le profil des champs récepteurs des cellules ganglionnaires X est *précisément* celui du laplacien d'une gaussienne (cf. figure 2).

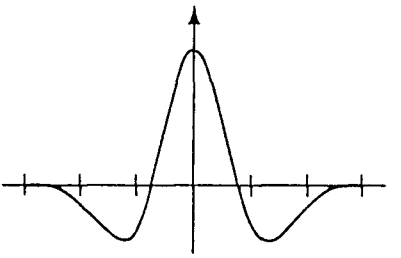

Figure 2. *Le « profil récepteur » du laplacien d'une gaussienne (d'après Marr, 1982, p. 55).*

Si une cellule X ON et une cellule X OFF voisines sont activées ensemble, cela détecte un « zero-crossing » et donc une discontinuité. Ce dispositif a été amélioré après Marr et a suscité de nombreux travaux et discussions (cf., par exemple, Haralick, 1984 ; Grimson-Hildreth, 1985 ; Richter-Ullman, 1986). Son accord avec l'expérience est remarquable.

On peut ainsi faire l'hypothèse qu'il existe des champs de cellules ganglionnaires dont la vocation fonctionnelle est la détection et l'explicitation de discontinuités qualitatives, *localement et à plusieurs échelles.* Ces champs ont une architecture uniforme et modulaire et ils calculent de façon massivement parallèle.

Il semble que l'algorithme de Marr soit un exemple, neurophysiologiquement implémenté, de ce que l'on appelle maintenant une analyse du signal par développement *en série d'ondelettes* (cf. Meyer, 1989). L'analyse en termes d'ondelettes est un processus multirésolution d'analyse de Fourier locale et multiéchelle qui consiste à développer une fonction f(x) (éventuellement très compliquée, fractale par exemple) appartenant à un certain espace fonctionnel (l'espace de Hilbert $L^2(\mathbb{R})$, par exemple) sur une base (orthonormée) d'ondelettes $\Psi_{j,k}$ construites à partir d'une seule fonction Ψ par dilatations et translations. On aura par exemple $\Psi_{j,k} = 2^{j/2}\Psi(2^j x\text{-}k)$ où $j,k \in Z$. Les coefficients $f_{j,k}$ du développement de f sur la base $(\Psi_{j,k})$ sont alors obtenus par convolution. Dans le dispositif de Marr, c'est ΔG — c'est-à-dire le profil d'un champ récepteur typique — qui joue le rôle d'ondelette.

3.3. La finalisation des algorithmes rétiniens par le problème inverse objectif.

Comme y insiste Y. Meyer, « une image contient une quantité énorme d'information et une grande partie de cette information est superflue ». Son analyse en termes d'ondelettes permet d'en extraire — d'en expli-

citer au sens de Marr — « diverses versions schématiques, simplifiées, dont le codage numérique et la transmission soient réalisables avec un coût raisonnable » (Meyer, 1989, p. 40). Il est remarquable que cette schématisation de l'information *coïncide avec une analyse morphologique objective de l'image.* La théorie mathématique des algorithmes de traitement de l'information et la vocation fonctionnelle de la base neurophysiologique implémentante rejoignent les contraintes et les finalisations imposées par les corrélats objectifs. Un « zero-crossing » stable à plusieurs échelles sera l'indice d'une discontinuité objective d'origine géométrique et physique. De telles discontinuités objectives seront préférentiellement traitées comme bords perceptuels. Dès les niveaux les plus précoces de la perception c'est donc son orientation vers les structures objectives (son intentionalité) qui domine. Et cette orientation n'est pas quelconque. Elle repose, insistons-y, sur une structuration *morphologique* du signal.

La base morphologique de la perception est donc imposée par la physiologie et les mathématiques. *Sa nécessité est d'origine à la fois informationnelle et objective. Une théorie mathématique morphologique doit donc être intégrée aux principes de la modélisation en perception visuelle.* C'est en particulier à partir d'elle — et cela pose un magnifique problème mathématique — qu'il faut retrouver les représentations symboliques opérant aux niveaux cognitifs supérieurs.

A propos de cette base morphologique, Marr remarque : « zero-crossing provides a natural way of moving from an analogue or continuous representation like the two-dimensional image intensity values I(x,y) to a discrete, symbolic representation » (p. 67). On ne saurait mieux exprimer le fait que le morphologique se situe entre le continu physique et le discret symbolique et que la vision *naturelle* le présuppose. Pour des systèmes naturels (où le discret symbolique ne peut pas exister d'emblée), *les discontinuités qualitatives morphologiques fournissent, une fois explicitées, la condition de possibilité de la constitution d'un niveau symbolique. En tant que singularités objectives encodées dans le signal, elles supportent l'information.*

> « The raw primal sketch is a very rich description of an image since it contains virtually all the information in the zero-crossings from several chanels. Its importance is that it is the first representation derived from an image whose primitives have a high probability of reflecting physical reality directly » (p. 71).

Comme l'explique T. Poggio :

> « Instead of raw numerical values of intensity, one seeks a more symbolic, compact and robust representation of the visual world : a description of the

world in which the primitive symbols — the signs in which the visual world is coded — are intensity variations » (Poggio, 1984, p. 72).

La structuration conceptuelle de l'image n'est donc pas, selon Marr, essentiellement descendante. Elle n'a pas à être entièrement inférée à partir de connaissances supplémentaires préalables. Elle est en grande partie reconstructible de façon ascendante à partir de la base morphologique extraite de ce que Marr appelle « the physics of the situation ». La connaissance supplémentaire nécessaire *n'est pas conceptuelle.* C'est une « general knowledge embeded in the early visual processes as general constraints, together with the geometrical consequences of the fact that the surfaces coexist in three-dimensional space » (p. 273).

4. L'esquisse 2-1/2D et le problème des contours apparents

Par globalisation, l'esquisse primaire « complète » explicite l'organisation morphologique de l'image. La question devient alors : *comment remonter de l'organisation morphologique 2D à des modèles 3D ?* Il est nécessaire de passer par un niveau intermédiaire et l'un des principaux mérites de Marr est d'avoir compris ce point fondamental.

4.1. Le problème du contour comme problème central de la vision computationnelle.

Marr appelle, nous l'avons vu, esquisse 2-1/2D le niveau de cette « intermediate vision » qui constitue le « pivotal point » de toute sa théorie. C'est le niveau de la « pure perception ». C'est « an internal representation of objective physical reality that *preceded* the decomposition of the scene into " objects " ». Pré-conceptuel, modulaire et ascendant, il représente et explicite « what the photons are carrying information about ». A ce titre, « it provides the cornerstone for an overall formulation of the entire vision problem » (p. 269-272).

Comme nous l'avons vu, l'esquisse 2-1/2 D intègre tout un ensemble de données issues des modules inférieurs et, en particulier, les données concernant les valeurs, les variations continues et les discontinuités de la profondeur (stéréopsie) et de l'orientation locale des surfaces. Énormément de travaux expérimentaux et mathématiques ont été consacrés à la façon dont les informations locales issues de la stéréopsie, de la texture, de l'ombrage, du mouvement et des contours coopèrent dans le processus de saisie perceptive d'une forme. Ce sont les problèmes « shape from stereo », « shape from texture », « shape from shading », etc. (cf., par exemple, Brady, 1982 ; Mingolla-Todd, 1986 ; Ikeuchi, 1984). Mais, selon nous, l'ensemble en est subordonné à la résolution d'un problème cen-

tral. En effet, d'après nos principes épistémologiques, les algorithmes de l'esquisse 2-1/2 D doivent être finalisés par le problème inverse objectif. Or, quelle est la nature de celui-ci à ce niveau ?

Le problème est le suivant. Comment remonter de distributions de discontinuités 2 D à des objets 3 D ? Cela n'est possible que si :

(i) on sait interpréter certaines discontinuités comme des *contours apparents* ;

(ii) on sait remonter des contours apparents d'un objet à cet objet lui-même.

Le premier problème est proprement perceptif. Il suppose que, au moyen des données de profondeur fournies par la stéréopsie ou des données de courbure et d'orientation de surfaces fournies par l'ombrage, etc., on puisse *désambiguïser* les multiples projections 3 D → 2 D pouvant aboutir à la même morphologie 2 D (entre deux domaines homogènes contigus séparés par un bord, lequel est devant et lequel est derrière ?, etc.).

Le second problème est en revanche *strictement géométrique et objectif.* Nous l'appellerons *le problème du contour* : comment est-il possible de reconstruire une forme géométrique 3 D à partir de ses contours apparents 2 D ? Ce problème est le problème central du niveau 2-1/2 D. C'est le noyau du problème inverse objectif car c'est sur lui que se concentre le *saut dimensionnel* 2 D → 3 D. *Sa résolution mathématique* devrait donc contraindre et finaliser de façon essentielle *l'ensemble des algorithmes 2-1/2 D de la vision computationnelle.* Or cela est très loin d'être le cas actuellement, la plupart des théoriciens de la vision ignorant les éléments de géométrie différentielle et de théorie des singularités exigés. Il faut dire que ceux-ci sont profonds et sophistiqués.

Encore une fois, Marr fait ici partiellement exception. A propos du saut dimensionnel, il remarque : « when one reflects upon it, this is actually quite an amazing fact » (p. 215). Et il pose bien le problème du contour comme problème central. Mais sa méconnaissance de certains récents résultats mathématiques puissants le conduit à faire des hypothèses *ad hoc.*

Soit T un objet (une forme) dans $\mathbb{R}^3$ et C son contour apparent (CA) relativement à une certaine projection Π.

(i) Marr introduit — et cela est correct — une hypothèse de *généricité* : T est en position générale par rapport à Π.

(ii) Il définit ensuite — et cela est également correct — le générateur du contour G, c'est-à-dire (cf. plus bas) le lieu critique de Π (qui est une courbe se projetant sur C).

(iii) Mais, comme il veut pouvoir reconstruire T à partir d'un seul CA et, pour ce faire, appliquer un théorème simple, il introduit l'hypothèse,

ad hoc et irréaliste, que le générateur G est *planaire* et que la forme T est un « cône généralisé », c'est-à-dire une surface engendrée en déplaçant une section variable le long d'une âme (cf. figure 3). Dans ce cas, en effet, C détermine bien T.

Cette hypothèse *ad hoc* est loin d'être innocente puisqu'elle conduit à décomposer les formes naturelles en cônes généralisés et, par conséquent, à imposer des contraintes non naturelles et non justifiées au niveau 3 D.

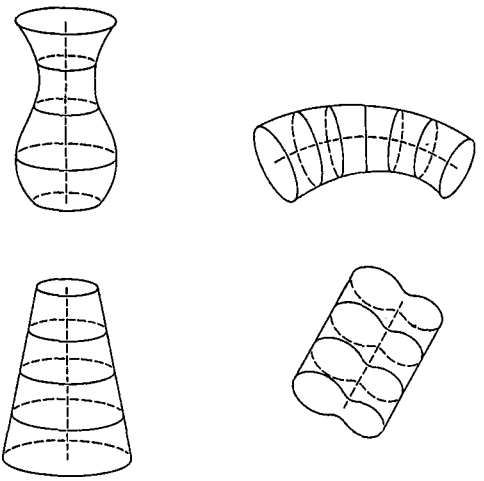

Figure 3. *Le concept de cône généralisé chez Marr (d'après Marr, 1982, p. 224).*

4.2. Le contenu géométrique du problème du contour.

Qu'est-ce que géométriquement le contour apparent (CA) d'un objet (d'une forme, d'une surface) T dans $\mathbb{R}^3$? Supposons pour fixer les idées que la surface T soit un tore (cf. figure 4). Se donner un CA de T consiste :

(i) à choisir dans $\mathbb{R}^3$ un plan de projection Δ ;

(ii) à choisir une direction de projection δ transverse à Δ ;

(iii) à considérer la projection Π de T sur Δ parallèlement à δ.

Le générateur du contour Γ est alors défini comme le *lieu critique* — ou le *lieu singulier* — de l'application $\Pi \mid_T : T \to \Delta$ restriction de Π à la surface T, c'est-à-dire comme le lieu des points $x \in T$ où la direction de projection δ est *tangente* à T. Le CA (géométrique) C est alors la projection Π (Γ) de Γ (en situation perceptive réelle, C ne sera en général que partiellement visible) (cf. figure 4).

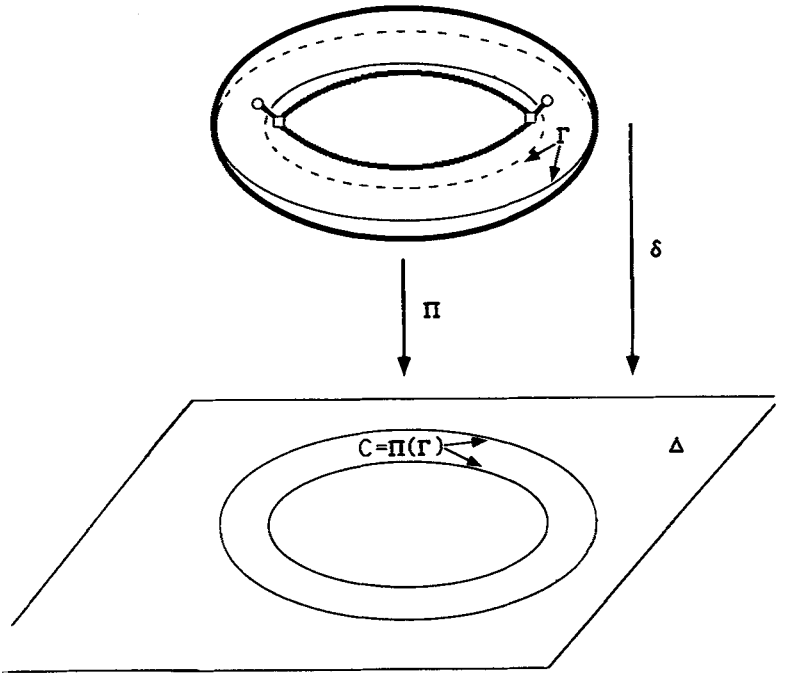

Figure 4. *Le contour apparent d'un tore T est l'ensemble des singularités de la projection* $\Pi : T \to \Delta$ *parallèlement à la direction* δ . Γ *est le générateur du contour et son image* $C = \Pi(\Gamma)$ *est le contour.*

On notera que le CA n'est pas seulement un ensemble C de courbes dans Δ. C'est un ensemble de courbes qui est un lieu critique, c'est-à-dire *un ensemble de singularités d'application.*

Le *type* de l'information morphologique que constitue un CA est donc loin d'être évident. Il reste incompréhensible en dehors de la théorie mathématique spécifique qui permet de le définir. *Les systèmes visuels sont des reconstructeurs de formes fondés sur des analyseurs de singularités eux-mêmes fondés sur des détecteurs de discontinuités.* Il s'agit là d'un fait fondamental.

Le programme de recherche d'une morphodynamique visuelle est donc bien défini.

(i) *Décrire et classer* les types de singularités locales pouvant (et devant) apparaître génériquement (stablement) dans les CA de surfaces.

(ii) Décrire et classer les singularités plus complexes pouvant (et devant) apparaître stablement dans des *déformations* génériques de CA.

(iii) Montrer qu'il existe *des formes normales algébriques* de ces singularités génériques. Cela est nécessaire pour pouvoir réduire la *géométrie* de celles-ci (qui, *a priori,* comprend une information infinie) *à une infor-*

mation numérique finie (pouvant donc être codée et transmise à un « coût raisonnable », cf. plus haut le problème analogue pour l'esquisse 2 D).

(iv) Reconstruire qualitativement *la géométrie différentielle* des surfaces à partir de la *famille* de leurs CA.

(v) Comprendre comment l'information morphologique 2-1/2 D *peut être encodée dans des champs 2 D de données numériques ponctuelles.* Cela est nécessaire à son calcul par des champs de processeurs ponctuels neuronalement implémentés.

(vi) Comprendre, enfin, comment les corrélats objectifs de cette information peuvent être encodés dans le signal lumineux. Une telle « optique morphologique » est nécessaire à la thèse réaliste, selon laquelle (a) il existe bien de tels corrélats objectifs et (b) l'esquisse 2-1/2 D explicite « ce sur quoi les photons véhiculent de l'information ».

Or, il se trouve qu'une partie considérable de ce programme de recherche est d'ores et déjà réalisée (pour des indications, cf. par exemple, Petitot, 1982, 1986b et, surtout, leurs bibliographies). Il paraît donc légitime, souhaitable et urgent d'intégrer tous ces résultats fondamentaux à la théorie de la vision computationnelle.

4.3. Le théorème de Whitney-Thom.

La surface T est une variété différentiable de dimension 2 plongée dans $\mathbb{R}^3$. On peut évidemment la décrire par ses équations. Mais une telle description est *extrinsèque*. Si l'on souhaite une description intrinsèque de sa géométrie, au niveau de structure différentiable, alors, comme le faisait déjà Gauss au début du siècle dernier, on doit introduire des *coordonnées locales*. En effet, si T est une surface *régulière* (sans singularités), elle est, en chaque point x, *localement* identifiable, au niveau de structure différentiable, à un plan. Une telle identification est réalisée par la donnée de coordonnées locales (x_1, x_2) en chaque point. Ces systèmes (dits cartes locales) se recollent entre eux à travers des changements différentiables de coordonnées locales[2]. Si (x_1, x_2) (resp. (y_1, y_2)) est un système de coordonnées locales au voisinage de x (resp. $\Pi(x)$)[3] l'application $f = \Pi \mid_T : T \rightarrow \Delta$ est localement décrite par un système d'équations $(y_1 = f_1(x_1, x_2); y_2 = f_2(x_1, x_2))$ où f_1 et f_2 sont des fonctions différentiables de deux variables réelles à valeurs réelles. $\Pi \mid_T$ est donc un cas particulier d'application différentiable $f : M \rightarrow N$ entre deux surfaces différentiables M et N ($M = T$ et $N = \Delta$).

Pour décrire *qualitativement* la géométrie locale de f, l'idée fondamentale est de généraliser le concept classique de série de Taylor, c'est-à-dire d'approximations successives de f par des applications *polynômiales*

2. Il est évidemment impossible de rappeler ici serait-ce des rudiments de géométrie différentielle.

3. Comme Δ est un plan, (y_1, y_2) est aussi un système de coordonnées global.

(donc algébriques) de degré de plus en plus grand. Cette méthode sera adéquate dans les cas — dits de *détermination finie* — où l'on saura démontrer que le développement de Taylor $T^k(f)$ de f à un ordre fini k suffit à caractériser qualitativement la géométrie locale de f (autrement dit que l'adjonction de termes d'ordre supérieur à k ne modifie $T^l(f)$ pour $l > k$ que quantitativement et non pas qualitativement, ou encore que l'écart entre $T^k(f)$ et $T^l(f)$ est résorbable par un changement approprié de coordonnées locales).

Le développement de Taylor à l'ordre 1 correspond à ce que l'on appelle *l'application linéaire tangente* D_xf de f en x. Elle représente la façon dont f agit infinitésimalement sur les vecteurs tangents à M en x. Ces vecteurs constituent un espace vectoriel T_xM (de dimension égale à celle de M) dit *espace tangent* à M en x. D_xf est une application linéaire $D_xf : T_xM \longrightarrow T_{f(x)}N$ qui est la meilleure approximation linéaire de f en x (comme la tangente à une courbe en un point est sa meilleure approximation linéaire en ce point). Relativement aux bases de T_xM et $T_{f(x)}N$ associées au choix de coordonnées locales (x_1, x_2) et (y_1, y_2), la matrice de D_xf est donnée par la matrice des dérivées partielles de f_1 et f_2, dite *matrice jacobienne* :

$$[D_xf] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}$$

La considération du type de D_xf permet déjà d'obtenir de précieux renseignements sur la géométrie locale de f en x. Soit $\mathfrak{M}$ l'espace vectoriel des matrices 2×2 $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ ($\dim \mathfrak{M} = 4$). Dans $\mathfrak{M}$, il existe une stratification naturelle *par le rang*, c'est-à-dire une décomposition Σ de $\mathfrak{M}$ en sous-variétés de dimensions décroissantes se recollant entre elles avec de bonnes propriétés d'incidence.

(i) La première strate est la strate Σ^0 des matrices régulières (c'est-à-dire des matrices de rang 2, des matrices de déterminant $|D| = ad\text{-}bc$ non nul, bref des matrices inversibles). Elle est topologiquement ouverte (donc de dimension 4) et dense dans $\mathfrak{M}$. En effet, une matrice $D \in \mathfrak{M}$ est en général de déterminant $|D| \neq 0$ et cette propriété est stable par petite perturbation des coefficients a, b, c, d.

(ii) La seconde strate Σ^1 est celle des matrices D de rang 1, c'est-à-dire des matrices (non identiquement nulles) de déterminant $|D| = 0$. C'est l'*hypersurface* de $\mathfrak{M}$ (moins l'origine) d'équation $ad\text{-}bc = 0$. Elle est donc de dimension 3, ou encore — si on appelle *codimension* d'une sous-variété W plongée dans une variété ambiante V la différence de dimensions $\dim V - \dim W$ — *de codimension 1.*

(iii) La troisième strate Σ^2 se réduit à l'origine. Elle ne contient que la matrice nulle de rang 0 D = 0.

On remarquera que la frontière de Σ^0 est $\Sigma^1 \cup \Sigma^2$ et que la frontière de Σ^1 est Σ^2.

L'application linéaire tangente D_xf détermine la structure locale de f en x au sens suivant. Notons $\Sigma^0(f)$, $\Sigma^1(f)$ et $\Sigma^2(f)$ les sous-ensembles de M constitués des $x \in M$ tels que $D_x(f) \in \Sigma^0, \Sigma^1, \Sigma^2$.

(i) Si $x \in \Sigma^0(f)$ (si $D_xf \in \Sigma^0$, c'est-à-dire si D_xf est inversible), l'application f est *localement inversible.* C'est un difféomorphisme local de (M, x) sur (N, f (x)) et sa géométrie est donc *qualitativement triviale.*

(ii) Si $x \in \Sigma^1(f)$ (c'est-à-dire si $D_xf \in \Sigma^1$), alors il existe une direction δ de T_xM qui se trouve *annulée* par D_xf. Autrement dit, le noyau $\mathrm{Ker}(D_xf)$ de D_xf n'est pas trivial (il n'est pas réduit à 0).

De façon générale, on dit que x est un *point critique* de f si $D_x(f)$ n'est pas de rang maximal. On dit alors que f (x) est une *valeur critique* de f. L'ensemble Σ (f) des points critiques de f est donc donné par $\Sigma(f) = \Sigma^1(f) \cup \Sigma^2(f)$. Nous allons voir que, sous certaines conditions, la géométrie locale reste déterminée à un ordre fini (et même à un ordre très bas). Mais notons d'abord que dans le cas particulier $f = \Pi \mid_T : T \to \Delta$, dire que $x \in \Sigma^1(f)$ revient à dire que la direction de projection δ appartient à T_xM et qu'elle est donc tangente à T. Autrement dit, *$\Sigma^1(f)$ n'est dans ce cas rien d'autre que le générateur du contour* Γ.

Il faut se convaincre que la complexité d'une application différentiable $f : M \to N$ peut être prodigieuse. Par exemple, on peut montrer (théorème dû à Borel) que si F est un fermé de M (et un tel F peut être d'une complexité infinie, fractale par exemple), il existe une application différentiable $f : M \to \mathbb{R}$ telle que $F = f^{-1}(0)$. Il est donc *impossible* de classer les types qualitatifs des f. Pour accéder malgré cela à une possibilité de classification, on applique la stratégie *de la stabilité structurelle.* Soit $\mathfrak{F}(M, N)$ l'espace fonctionnel des f. Sur $\mathfrak{F}$ il existe une topologie $\mathfrak{T}$ (dite topologie de Whitney) naturellement adaptée au niveau de structure différentiable (intuitivement, c'est la topologie de la convergence uniforme des fonctions et de toutes leurs dérivées partielles sur les compacts de M, avec en plus une contrainte d'identité « à l'infini », c'est-à-dire sur le filtre des complémentaires des compacts). D'autre part, sur M et sur N il existe les changements de coordonnées *globaux* que sont les difféomorphismes. Il est évident que deux applications $f, g \in \mathfrak{F}$ sont *qualitativement (différentiablement) équivalentes,* si elles sont conjuguées par de tels difféomorphismes, autrement dit s'il existe $\varphi \in \mathrm{Diff}(M)$ et $\psi \in \mathrm{Diff}(N)$ tels que $g = \psi_o f_o \varphi^{-1}$. Autrement dit, le groupe $G = \mathrm{Diff}M \times \mathrm{Diff}N$ opère sur F *comme un groupe de relativité* et les orbites de son action sont *les types qualitatifs* des éléments $f \in \mathfrak{F}$.

On dit alors que f est *structurellement stable* s'il existe un voisinage de

f (pour la topologie $\mathfrak{T}$) dont tous les éléments g sont qualitativement équivalents à f (autrement dit, si le type qualitatif de f résiste aux petites perturbations). La stratégie de la stabilité structurelle, introduite par Whitney en 1955 et considérablement développée par Thom, Arnold et d'autres, consiste :

(i) à analyser d'abord la géométrie locale des applications *structurellement stables* ;

(ii) à analyser ensuite celle des applications instables, mais en introduisant *progressivement* des degrés de plus en plus grands d'instabilité (cela suppose évidemment que l'on ait explicité les causes possibles d'instabilité structurelle).

La structure locale des applications structurellement stables entre surfaces est entièrement connue. Elle est résumée dans le théorème suivant.

Théorème de Whitney-Thom :

1. Les applications structurellement stables $f : M \longrightarrow N$ sont *génériques* dans $\mathfrak{F}$: toute application $g \in \mathfrak{F}$ est approximable aussi près que l'on veut par une application f structurellement stable.
2. Si f est structurellement stable, sa géométrie locale est équivalente à celle de l'un des trois modèles locaux algébriques suivants :

(a) $y_1 = x_1$, $y_2 = x_2$: *point régulier* (f est un difféomorphisme local),
(b) $y_1 = x_1^2$, $y_2 = x_2$: *point pli,*
(c) $y_1 = x_1^3 + x_1x_2$, $y_2 = x_2$: *point fronce.*

Ce théorème montre que, sous l'hypothèse de stabilité structurelle, la géométrie locale de f est *déterminée à l'ordre 2* (c'est-à-dire par T^2 (f)). Ce résultat est fondamental pour ce qui nous occupe ici puisqu'il ne s'agit de rien de moins que d'un théorème *de réduction d'une information morphologique à une information algébrique finie.* Il existe des *modèles locaux algébriques universels* pour la géométrie locale. En hommage à Marr, nous les appellerons des modèles 2-1/2 D.

La structure géométrique d'un pli est évidente (cf. figure 5).

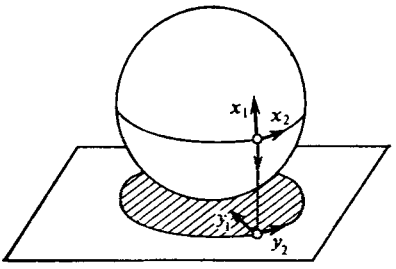

Figure 5. *La structure d'un point pli (d'après Arnold, 1986, p. 16).*

Celle d'une fronce est un peu plus complexe. Mais elle est facile à dériver de son modèle algébrique. Pour la visualiser, considérons le graphe de $y_1 = x_1^3 + x_1x_2$ dans l'espace $\mathbb{R}^3$ de coordonnées (x_1, $x_2 = y_2$, y_1). f est la projection de ce graphe sur le plan (y_1, y_2) (cf. figure 6).

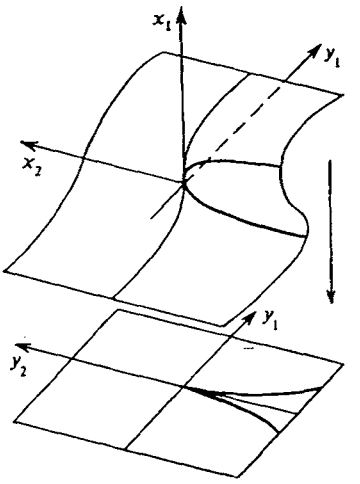

Figure 6. *La structure d'un point fronce (d'après Arnold, 1986, p. 16).*

La matrice jacobienne de f en x est $[D_xf] = \begin{bmatrix} 3x_1^2 + x_2 & x_1 \\ 0 & 1 \end{bmatrix}$. Donc $x \in \Sigma^1(f)$ si $D_xf = 3x_1^2 + x_2 = 0$ (équation d'une parabole Γ dans le plan (x_1, x_2) : le générateur du contour). $C = f(\Gamma)$ est donc la parabole semi-cubique du plan (y_1, y_2) d'équations paramétriques $y_1 = -2x_1^3$, $y_2 = -3x_1^2$ (car $x_2 = -3x_1^2$ sur Γ). Elle possède à l'origine un point de rebroussement appelé un point *cusp*. On voit qu'aux points plis de Γ, le noyau $\mathrm{Ker}D_xf$ (la direction de projection) est *transverse* à Γ. En revanche au point fronce $x = 0$, $\mathrm{Ker}D_xf$ est au contraire *tangente* à Γ.

Autrement dit, en un point pli x, f est singulière, $x \in \Sigma^1(f)$, mais la restriction $f \mid \Sigma^1(f)$ de f à $\Sigma^1(f) = \Gamma$ est, elle, *régulière*. C'est pourquoi il est naturel de noter $\Sigma^{1,0}(f)$ l'ensemble des points plis de f. En revanche en un point fronce x de f, $x \in \Sigma^1(f)$ mais la restriction $f \mid \Sigma^1(f)$ est *singulière*. On note donc $\Sigma^{1,1}(f)$ l'ensemble des points fronces de f.

On remarquera d'autre part que, dans ces modèles locaux, $\Sigma^2(f) = \varnothing$. Nous allons voir qu'il s'agit là d'une *nécessité* sous l'hypothèse de stabilité structurelle.

On voit ainsi apparaître l'idée fondamentale de *types* de points singuliers *finiment* descriptibles et d'une *hiérarchie* de ces types. Les lieux singuliers des applications stables sont des *stratifications*, des « empilements » de lieux singuliers de restrictions à des lieux singuliers : Sing(f),

Sing(f | Sing(f)), Sing(f | Sing(f | Sing(f))), etc., la stabilité structurelle *bornant* ce type d'itération par les *dimensions* de M et de N.

4.4. La théorie des jets et le processing ponctuel des géométries locales.

Bien que les termes *homogènes* du développement de Taylor d'une application ne possèdent pas de *signification géométrique intrinsèque* (indépendante du choix, conventionnel, des coordonnées locales), on peut montrer que les développements $T^k(f)$ jusqu'à un rang donné possèdent, eux, une signification géométrique intrinsèque. Cela a permis à Charles Ehresmann d'élaborer *la théorie des jets* qui fournit une réponse au problème fondamental *de l'encodage d'une géométrie locale par un champ de données numériques ponctuelles* (problème 4.2.(v)).

L'idée généralise cette opération bien connue qui consiste, étant donné une courbe $y = f(x)$, à considérer le *champ* de ses tangentes T_x en chaque point et à reconstruire la courbe comme *enveloppe* de ses tangentes.

Soit $f : M \to N$. En chaque point $x \in M$, le développement de Taylor au premier ordre est constitué de trois groupes de données *ponctuelles* :

(i) $x \in M$: deux coordonnées : x_1, x_2 ;

(ii) $y = f(x) \in N$: deux coordonnées : $y_1 = f_1(x_1, x_2)$, $y_2 = f_2(x_1, x_2)$;

(iii) $D_x f \in \mathfrak{D}$: quatre coordonnées : $a = \frac{\partial f_1}{\partial x_1}(x_1, x_2)$, $b = \frac{\partial f_1}{\partial x_2}(x_1, x_2)$, $c = \frac{\partial f_2}{\partial x_1}(x_1, x_2)$, $d = \frac{\partial f_2}{\partial x_2}(x_1, x_2)$.

Ces *huit* données numériques constituent le *1-jet* de f en x, 1-jet noté $j^1f(x)$. $j^1f(x)$ habite naturellement dans un espace à huit dimensions qui, localement, est le produit direct $M \times N \times \mathfrak{D}$. Lorsque l'on globalise, ces produits directs se recollent en un fibré vectoriel de base $M \times N$, appelé espace (ou fibré) des 1-jets des applications différentiables $f : M \to N$ et noté $J^1(M, N)$. Si $f \in \mathfrak{F}$, on lui associe son 1-jet j^1f qui est l'application de M dans $J^1(M, N)$ définie par le *champ* des 1-jets $j^1f(x)$:

$$j^1f : M \to J^1(M, N)$$
$$x \to j^1f(x).$$

Mais nous avons vu que, dans les fibres $\mathfrak{D}$ de $J^1(M, N)$ — et donc dans $J^1(M, N)$ — il existe une stratification naturelle $\Sigma = (\Sigma^0, \Sigma^1, \Sigma^2)$. Il est clair que l'on a $\Sigma^0(f) = (j^1f)^{-1}(\Sigma^0)$, $\Sigma^1(f) = (j^1f)^{-1}(\Sigma^1)$, $\Sigma^2(f) = (j^1f)^{-1}(\Sigma^2)$. *La stratification $\Sigma(f)$ de la source M opérée par f au moyen du rang de l'application linéaire tangente $D_x f$ n'est donc rien d'autre que l'image réciproque de la stratification universelle Σ par le 1-jet $j^1(f)$ de f.*

Le théorème de Whitney montre que les 2-jets $j^2(f)$ *suffisent* pour reconstruire qualitativement la géométrie locale de *toutes* les applications *structurellement stables*. En généralisant aux 2-jets (cela est trop technique pour être exposé ici) les constructions précédentes, on en arrive à

la conclusion que, génériquement, la géométrie locale de f est descriptible *à partir de l'image inverse, par les 2-jets $j^2(f)$, de stratifications universelles* (indépendantes de f) des espaces de jets. Autrement dit, *la géométrie locale est calculable au moyen des champs de données numériques que sont les jets d'ordre $\leqslant$ 2. La théorie des jets est donc bien fondamentale pour la vision computationnelle puisqu'elle explique comment des champs de processeurs ponctuels possédant une bonne rétinotopie peuvent calculer de la géométrie, c'est-à-dire traiter de l'information morphologique.*

Il est d'ailleurs étrange que les spécialistes de la vision aient été aussi peu attentifs jusqu'ici à l'une des idées les plus profondes et les plus fécondes de toutes les sciences, à savoir celle de la dialectique du local et du global. L'idée est que les contraintes (lois de la nature, etc.) se décrivent au niveau local par des équations sur des jets et que, par intégration, elles admettent pour solutions des entités globales. Par exemple, une équation différentielle ordinaire consiste à se donner en chaque point d'un espace M (espace de configurations ou espace de phases d'un système mécanique, etc.) un vecteur tangent $X(x) \in T_xM$ et à chercher les trajectoires intégrales d'un tel champ. De même, un feuilletage (un système de Pfaff) consiste à se donner en chaque point $x \in M$ un sous-espace vectoriel $P(x)$ de T_xM et à chercher les variétés intégrales. De même encore, une équation aux dérivées partielles est une équation dans un espace de jets convenable. Par exemple, une équation de diffusion (à une dimension) comme $\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2}$ s'exprime par l'équation $a = m$ dans l'espace de jets $J^2(\mathbb{R}, \mathbb{R})$ de coordonnées $(x ; y = f(x) ; a = \frac{\partial f}{\partial t}, b = \frac{\partial f}{\partial x} ; k = \frac{\partial^2 f}{\partial t^2}, l = \frac{\partial^2 f}{\partial x \partial t}, m = \frac{\partial^2 f}{\partial x^2})$.

L'importance des théories évoquées plus haut est d'avoir montré que l'analyse morphologique peut se ramener à de tels champs de données numériques, champs dont les formes sont en quelque sorte des solutions intégrales. Elle donne une nouvelle dimension à l'intuition initiale des Gestalt-théoriciens.

Certes, de très nombreux modèles de vision computationnelle consistent à reconstruire des formes à partir de champs de données locales. Par exemple, on cherchera à associer à chaque point de l'image une orientation locale de surface (ce qui est équivalent à une direction normale : on cherche à reconstruire l'application de Gauss de la surface, cf. p. 172) obtenue à partir des informations sur la stéréopsie, l'ombrage, le gradient, la texture, etc. (cf., par exemple, les articles, déjà cités, Brady, 1982 ; Ikeuchi, 1984 ; Mingolla-Todd, 1986). Mais de tels modèles restent très en deçà des ressources actuelles de la géométrie différentielle.

Revenons à la théorie des jets. D'après un théorème fondamental de Thom, dit *théorème de transversalité,* la stabilité structurelle s'exprime par

des propriétés de transversalité des applications jets $j^k(f)$ sur les stratifications universelles des espaces de jets $J^k(M, N)$. Cela implique que, lorsque l'on prend les images réciproques de ces stratifications, leur structure géométrique soit préservée autant qu'il est possible. Cela implique à son tour *une borne drastique à la complexité* des singularités génériques. Considérons, par exemple, la strate Σ^2 de $J^1(M, N)$. Elle est de codimension 4. Comme dim $M = 2$, l'image $j^1f(M)$ de M dans $J^1(M, N)$ par j^1f est (au plus) de dimension 2. La stabilité implique la transversalité, et la transversalité implique à son tour, pour une simple raison de dimension ($2 < 4$), que $j^1f(M)$ *évite* Σ^2. C'est pourquoi, si f est structurellement stable, on a nécessairement $\Sigma^2(f) = \varnothing$. On montre aussi que, sous la même hypothèse, $\Sigma^1(f) = \Gamma = (j^1f)^{-1}(\Sigma^1)$ est une courbe régulière de M.

4.5. La solution du problème inverse objectif.

Sur le plan global, on peut montrer que si $f : M \to N$ est structurellement stable et si M est compacte, alors le générateur Γ est une courbe régulière et le CA $C = f(\Gamma)$ ne peut présenter comme singularités que des cusps isolés et des croisements normaux. Il existe alors des relations précises entre le nombre de cusps et la structure globale de M et de N (par exemple, leur caractéristique d'Euler-Poincaré).

D'autre part, on peut aussi classifier et mettre sous forme normale les singularités qui apparaissent stablement lors de *déformations* génériques de CA. La plus importante est *la queue d'aronde* où un point pli devient instable et se stabilise en engendrant deux cusps (cf. figure 7 pour l'exemple du tore). Le nombre de types qualitatifs de CA que peut présenter une forme T, ainsi que leurs relations d'incidence, fournit un renseignement fondamental sur *la complexité morphologique* de T.

L'ensemble de ces résultats (que nous n'avons fait qu'esquisser de façon très élémentaire) permet de résoudre le problème inverse *objectif*. Celui-ci impose à la résolution computationnelle du problème inverse *cognitif* les contraintes suivantes.

(i) Il doit exister des dispositifs de détection et de représentation (d'explicitation) des lignes de discontinuités (projections de points plis), des points d'arrêt de telles lignes (points cusps dont en général une des branches de points plis sera occultée si la surface est opaque), et des croisements de telles lignes (en général, une partie de la ligne pli arrière sera occultée et le croisement sera donc en forme de T). *L'algorithme de Marr correspond au premier de ces dispositifs.* Il faut donc en généraliser le principe aux deux autres cas.

(ii) Il faut pouvoir associer à ces primitives morphologiques 2 D (plis,

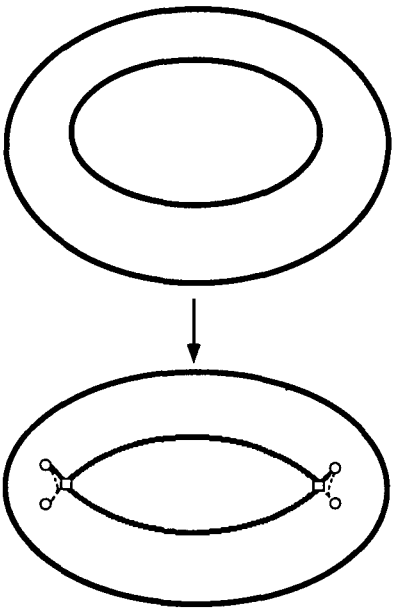

Figure 7. *Singularité de transition « queue d'aronde » pouvant apparaître stablement dans une déformation générique de contour apparent. Un point pli dégénère et engendre deux cusps (petits cercles) et un croisement (petit carré).*

cusps, croisements) les modèles locaux 2-1/2 D correspondants. Nous avons vu comment cela était possible.

5. Les travaux de Jan Koenderink

Jan Koenderink est l'un des rares spécialistes de la vision qui ait compris tout le bénéfice que la vision computationnelle peut tirer de l'usage des théories mathématiques évoquées plus haut pour résoudre le problème du « jump between logical levels (i.e. from the physical to the semantic domain) » (Koenderink, 1987, p. 367). Dans une série d'articles remarquables, parus pour la plupart dans *Biological Cybernetics*, il les a appliquées à tout un ensemble de problèmes[4].

5.1. Le point de vue épistémologique.

Adoptant une perspective « écologiste », Koenderink considérait dès 1976 :

4. Je remercie S. Thorpe de m'avoir récemment signalé ces travaux, apparemment inconnus jusqu'ici dans les milieux mathématiques pourtant directement concernés.

« that prior to the study of visual shape perception or visual egocentric localization an inventory of the invariants of the optical input under voluntary deplacements of the observer, has to be made. Such invariants pertain to objective geometrical properties of the environment. »

Mais il ajoutait aussitôt : « However, a comprehensive quantitative theory of the geometrical invariants of optical stimulation does not exist » (Koenderink, 1976, p. 51). Il introduisait alors l'idée directrice que l'information pertinente est concentrée dans les singularités des projections visuelles et que c'est donc la théorie des singularités qui permet de fonder mathématiquement un « écologisme » scientifiquement légitime.

5.2. La résolution du problème du contour.

Une des premières réussites de J. Koenderink est d'avoir explicitement utilisé dans sa théorie la résolution du problème du contour. Soit encore une fois notre forme T plongée dans $\mathbb{R}^3$. Ce que nous avons dit reste essentiellement valable si, au lieu de considérer une projection parallèle Π, nous considérons la projection Π_p de T sur Δ à partir d'un point de vue p extérieur à T. Soit $F_p : T \rightarrow \mathbb{R}$ la fonction *distance* d(p, x) de p à $x \in T$. Les points critiques de F_p sont ceux pour lesquels d(p, x) est stationnaire, c'est-à-dire ceux pour lesquels la direction px est orthogonale à T. Génériquement, ce sont des minima, des maxima et des cols. Soit w(x) le vecteur unitaire de la direction px d'origine x. Soit $v(x) \in T_xT$ sa composante tangentielle : v(x) définit un champ de vecteurs tangents sur T dont les trajectoires sont les lignes de pente de la distance. Les points critiques de F_p sont les points critiques ($v(x) = 0$) de ce champ.

En couplant ce champ au CA, on obtient ce que Koenderink appelle *un aspect.* « The aspect is a Gestalt-like feature of the visual input. » Il détermine le pattern d'excitation corticale. Il est constitué des CA orientés, des occultations de bords, des cusps, des croisements de lignes pli, des points où les lignes de pente de F_p touchent un bord occluant, des points critiques de F_p avec leur type (minimum, maximum, col), des séparatrices des directions de lignes de pente, des lignes de pente issues des cusps et des lignes de pente touchant un bord occluant (cf. figure 8).

La considération des aspects et de leurs déformations lors des changements de points de vue ou des déplacements d'objets permet non seulement de reconstruire la topologie de la surface T et sa structure différentiable, mais également de reconstruire partiellement ses propriétés riemaniennes (donc métriques). Cela signifie la chose suivante. On sait (depuis Gauss) que si l'on considère une surface plongée dans $\mathbb{R}^3$ comme une variété riemanienne, sa structure métrique est localement elliptique, hyperbolique ou parabolique. En un point hyperbolique, il

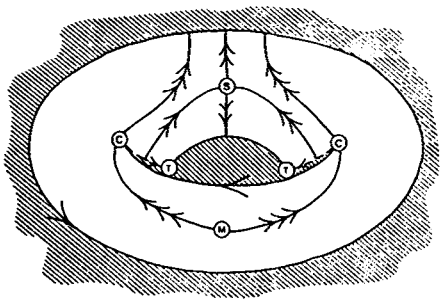

Figure 8. *Le concept d'aspect chez Koenderink.* → *: contour;* →→ *: séparatrice;* →→→ *: chemin passant par un cusp; C : point cusp; M : minimum de la fonction distance; S : point col; T : croisement normal (d'après Koenderink, 1979, p. 214).*

existe deux directions principales. Les trajectoires de ces deux champs de directions (dites lignes asymptotiques) admettent pour enveloppe les lignes de points paraboliques.

Koenderink montre que la famille des CA de T permet de déterminer le *type* de la métrique en chaque point (et donc en particulier les propriétés de convexité de T)[5]. Pour cela, il analyse avec soin les composantes de CA introduites par les déplacements du point de vue p sur T et il en explicite la structure à partir de *l'application de Gauss* de T, c'est-à-dire de l'application $G : T \rightarrow S^2$ qui à $x \in T$ associe le vecteur normal unitaire (externe, on suppose T orientable) n(x) à T en x (S^2 est la sphère unité de $\mathbb{R}^3$)[6]. Les lignes paraboliques de T correspondent aux plis P_G de G. Si x est un point du générateur Γ d'un CA de T, n(x) est normal à la direction δ (on suppose p à l'infini pour simplifier). L'image de Γ par G est donc incluse *dans un grand cercle* Γ_G de S^2. Lorsque l'on bouge δ, Γ_G se déplace et, en étudiant les transformations de sa position par rapport à P_G, on peut reconstruire qualitativement la géométrie proto-riemanienne de T. Cela résout le problème du contour.

A partir de cet acquis, Koenderink développe alors l'argument suivant, qui nous paraît fondamental. La déformation, par transformation des positions relatives de p et de T, des CA de T — qui ont une réalité perceptive bien établie — permet de reconstruire la géométrie *intrinsèque* (objective) de T. Elle permet donc de *prédire — d'anticiper sur —* ces déformations. Celles-ci, parce que prédictibles, peuvent être interprétées

5. Ce niveau plus fort que le différentiable et plus faible que le métrique ne semble pas avoir été très étudié mathématiquement. Il est en quelque sorte encore qualitatif et déjà proto-métrique, bien que sans notion de distance et de géodésiques.

6. L'application de Gauss est évidemment couramment utilisée dans les modèles de vision computationnelle puisqu'elle représente le champ des orientations locales d'une surface (cf. plus haut). Mais en général on n'utilise pas sa relation avec les CA.

comme d'origine *proprioceptive,* ce qui explique *l'invariance* objective de l'objet malgré la grande variation subjective de l'*input* visuel. « Our geometrical theory enables us to understand the structure of the observer's internal models of external bodies » (Koenderink, 1976, p. 59).

Dans un travail plus récent, Koenderink aborde la généralisation de la théorie de Marr. Sa première idée est d'abord, étant donné un pattern d'intensité 2D I(x, y), d'en représenter la morphogenèse en l'incluant dans une déformation $F = I_t$ conduisant de $I = I_1$ à un pattern I_0 trivial. La déformation inverse $I_0 \rightarrow I_1$ est donc un chemin de genèse de I. Koenderink choisit alors pour déformation une solution d'une *équation de diffusion* (type équation de la chaleur) $\frac{\partial F}{\partial t} = \Delta F$. La raison en est qu'une telle solution équivaut à lisser I par convolution avec une gaussienne (dépendant de t). On reprend donc l'algorithme de Marr mais en lui donnant un nouvel éclairage : « Gaussian blurring is the only sensible way to embed a primal image into a one-parameter family » (Koenderink, 1979, p. 365). L'auteur étudie ensuite la structure locale de F en termes

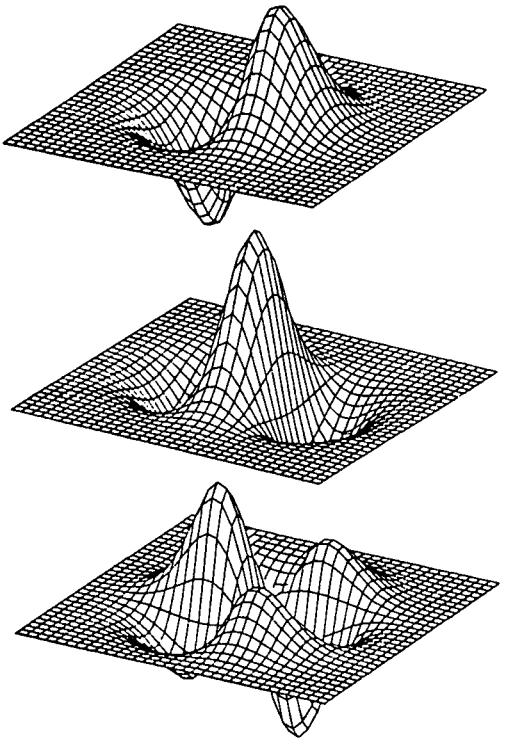

Figure 9. *Profils récepteurs permettant selon Koenderink d'effectuer des calculs de jets (d'après Koenderink, 1987, p. 371).*

de jets. Pour cela, il reprend l'idée directrice de Marr d'une analyse multirésolution (locale et multiéchelle) *par des convolutions avec des profils bien choisis f_n (x, t) de champs récepteurs.* Ces f_n (x, t) sont, comme chez Marr, des dérivées partielles (d'ordre n) de gaussiennes (cf. figure 9).

Koenderink explique alors l'importance du concept de jet pour la vision computationnelle. Dans F, l'information morphologique est continuement distribuée. Elle est *multilocale.* Dans les jets j^kF elle devient au contraire *ponctuelle* et traitable par des processeurs ponctuels.

> « Routines accessing a single location may aptly be called point processors, those accessing multiple location *array processors.* The difference is crucial in the sense that point processors need no geometrical expertise at all, whereas array processors do » (Koenderink, 1987, p. 370).

Les profils de champs récepteurs fournissent une implémentation des détecteurs de données différentielles. A partir d'eux, on peut construire *des processeurs de jets* qui sont des *détecteurs de traits morphologiquement significatifs.* « The order of the jets in the representation determines the " features " (the geometrical properties) that can be computed by a point processor » (*ibid.,* p. 370). Certaines hypercolonnes corticales seraient des champs de tels détecteurs :

> « the modules (like " cortical columns " in the physiological domain or " records " of raw data in the syntactic domain) of the sensorium are local approximation (N^{th} order jets) of the retinal illuminance that can be adressed as *a single datum* by the point processors. »

Les jets sont des K-uples de *nombres* possédant « a semantic content in terms of certain visual routines ».

> « That looking at a retinal illuminance distribution through a receptive field profile (or even through several layers of them !) is equivalent to looking at certain partial derivative of a blurred pattern is a new insight that immediately leads to useful interpretation in terms of differential geometry » (*ibid.,* p. 374).

6. Éléments d'une théorie morphodynamique intégrée

Dans ce qui précède, nous nous sommes focalisés sur *un* point qui nous paraissait névralgique. Nous aimerions maintenant brièvement faire le lien avec d'autres recherches qui sont susceptibles de conduire à une théorie intégrée.

6.1. La nature d'une optique morphologique.

Une thèse réaliste sur l'information morphologique n'est évidemment tenable que si l'on peut montrer que les CA (c'est-à-dire des singularités d'applications différentiables) peuvent effectivement être encodés dans le signal lumineux, c'est-à-dire dans des solutions des équations de Maxwell. Le problème est loin d'être trivial mathématiquement. Il est résolu (depuis peu de temps seulement) en ce qui concerne un cas plus simple que celui des CA, à savoir celui des caustiques. Les caustiques sont les enveloppes de rayons lumineux qui apparaissent lorsque des faisceaux lumineux sont soumis à des contraintes dioptriques de convergence (de focalisation). Ce sont des singularités (des lieux critiques d'applications) faciles à décrire géométriquement. Elles dominent les images optiques et sont phénoménologiquement structurantes. Comment sont-elles encodées dans le signal optique ? Comment leur information, typiquement de nature morphologique, peut-elle être, comme dirait Marr, véhiculée par les photons ? Une réponse peut être donnée dans le cadre de l'approximation géométrique de l'équation des ondes. On montre (c'est très technique : théorie des intégrales oscillantes) qu'à chaque singularité générique de caustique (pli, cusp, ombilic, etc.) est associée une intégrale oscillante typique qui est une structure ondulatoire fine construite sur l'infrastructure géométrique de la singularité (pour des précisions, cf. Arnold *et al.*, 1986 ; pour une introduction, cf. Petitot, 1986b, 1989g ainsi que leurs bibliographies). Il s'agit là d'un exemple, en tous points remarquable et entièrement mathématisé, d'émergence qualitative de formes perceptivement significatives à partir de la physique fondamentale.

Contrairement aux idées reçues, il existe donc bien une optique morphologique et il est par conséquent légitime de faire l'hypothèse que l'information géométrique est non seulement géométriquement objective, mais également *physiquement* objective.

6.2. Le niveau 2-1/2 D, le niveau 3 D et la Structure conceptuelle.

Le niveau 3 D objective le niveau morphologique 2-1/2 D. Ses algorithmes commencent à être bien compris (on a, par exemple, étudié en détail le nombre minimal de projections planes et de CA dont on a besoin pour reconstruire de façon non ambiguë une forme tridimensionnelle : cf. entre autres Hoffman-Bennett, 1986). Mais beaucoup de ses constituants s'enracinent dans le niveau 2-1/2 D. Par exemple, la décomposition (relativement canonique) d'un objet en parties s'opère essentiel-

lement sur des bases morphologiques (les lignes de décomposition sont des lignes de forte courbure, etc.). Contrairement à ce que l'on croit habituellement, elle n'est pas descendante (reconnaissance d'occurrences de modèles de parties prototypiques stockées dans une mémoire à long terme) mais ascendante. Ainsi que l'affirment Hoffman et Richards :

> « the visual system decomposes shapes into parts [...] using a rule defining part boundaries rather than part shapes, [...] the rule exploits a uniformity of nature — transversality, and [...] parts with their description and spatial relations provide a first index into a memory of shapes » (Hoffman-Richards, 1984, p. 65).

Ce n'est que postérieurement à cette décomposition morphologique qu'interviennent les segmentations en constituants géométriquement typiques reposant sur un vocabulaire fini de primitives (cf., par exemple, Biederman, 1987) et que les niveaux supérieurs de représentation et d'organisation hiérarchisée de l'information visuelle deviennent de format similaire à, et compatibles avec, ceux de l'information non visuelle.

De même les phénomènes de *catégorisation* proviennent essentiellement de la forte non-linéarité du contrôle des formes par des paramètres de déformation (cf. Petitot, 1989e). A l'intérieur des catégories, les formes sont stables par rapport à la variation du contrôle. Les frontières des catégories sont, au contraire, des lieux critiques à la traversée desquels les formes deviennent structurellement instables par rapport au contrôle et, donc, changent de type qualitatif. De façon générale, ainsi qu'y insiste R. Jackendoff, énormément de traits qui servent à catégoriser les objets *sont morphologiques et non pas sémantiques.* Le niveau 3D est celui où le langage se branche sur la vision à travers la structure conceptuelle et le langage en hérite *de fortes composantes morphologiques* (au sens adopté ici, non linguistique, de morphologie).

6.3. Vision et langage.

Dans un certain nombre de travaux (en particulier, Petitot, 1979, 1982, 1985, 1989a, c, f) nous avons développé l'idée maîtresse de Thom selon laquelle les relations actantielles entre les actants spatio-temporels d'une scène visuelle étaient morphodynamiquement — et non pas seulement symboliquement — descriptibles.

Nous avons montré comment cette idée permettait de fonder et de développer mathématiquement ce que l'on appelle *l'hypothèse localiste* en linguistique et d'en déduire une théorie actantielle (casuelle), une théorie de l'aspectualité et une théorie de l'agentialité. Nous avons, enfin,

analysé le rapport qu'une telle schématisation morphodynamique entretient avec certains des courants fondamentaux de la linguistique cognitive actuelle (Langacker, Talmy, Jackendoff).

Enfin nous avons montré comment une telle théorie de la syntaxe actantielle permettait de répondre aux objections de principe élevées par J. Fodor et Z. Pylyshyn contre le connexionnisme (cf. Petitot, 1989f, i).

III. — LA MORPHODYNAMIQUE VISUELLE COMME RÉPONSE AUX PROBLÈMES DE L'ÉCOLOGISME ET DE L'INTENTIONALITÉ

Sur le plan *épistémologique,* nous considérons que l'existence d'un niveau de réalité morphologique assurant la médiation entre le physique et le symbolique est d'une grande importance dans la mesure où elle permet de résoudre un certain nombre de problèmes cruciaux qui resteraient autrement aporétiques. Donnons brièvement, pour conclure, quelques indications à propos de deux d'entre eux.

1. L'objectivité écologique

Dans un important article, J. Fodor et Z. Pylyshyn ont ruiné théoriquement les thèses écologistes. Ils partent de l'hypothèse classique : parce que cognitive, la perception doit nécessairement être un processus computationnel symbolique et inférentiel. Ils cherchent alors à invalider la thèse gibsonienne selon laquelle la perception est à même d'extraire de l'environnement des invariants possédant un contenu objectif. Pour cela ils dégagent, avec une acuité remarquable, les inconsistances de la théorie écologique. Selon eux, la principale consiste à fonder toute la théorie sur l'existence d'une information objective (mais non physique) qui serait présente dans le médium lumineux (discontinuités, déformations, formes, textures, réflectances, etc. des surfaces visibles), alors qu'on reste dans l'impossibilité de la définir. Que peut être, en effet, cette énigmatique « information in the light » (Fodor-Pylyshyn, 1981, p. 143) ? Pour les auteurs, en vertu du dualisme physique/symbolique, l'information est soit physique, soit symbolique. Si donc elle n'est pas à proprement parler physique, mais « écologique », elle doit nécessairement être symbolique. Bref, Gibson introduit une objectivité écologique introuvable. Il critique d'un côté la physique physicaliste et de l'autre la psychologie mentaliste. Il ne fournit toutefois pas d'alternative. D'où un cercle vicieux. « What we

need, of course, is some criterion for being ecological *other than perceptibility*. This however, Gibson fails to provide » (p. 146). Il faudrait une optique écologique différente de l'optique physique, capable de caractériser ce qui est *phénoménologiquement significatif*. Or, celle-ci demeure, selon les auteurs, inaccessible.

Le syllogisme est au fond le suivant. La seule extraction directe d'invariants ne peut être que celle effectuée par la transduction. Les transducteurs ne peuvent être sensibles qu'aux propriétés physiques du signal lumineux car leur fonctionnement est régi par des lois et les seules lois existantes sont les lois physiques. Il ne saurait donc exister de transducteurs (même compilés, c'est-à-dire opérant modulairement jusqu'à des niveaux post-rétiniens) qui extraient du signal des propriétés écologiques non physiques. Fodor et Pylyshyn dénoncent alors ce qu'ils considèrent être une *subreption* chez Gibson. Pour Gibson, il existe *de l'information contenue dans la lumière*. Mais, selon les auteurs, le concept d'information est *relationnel. La lumière contient de l'information sur l'environnement*, et « contenir de l'information sur » signifie « être corrélé avec ». Les propriétés de l'environnement sont donc *inférées* à partir de la structure du signal lumineux sur la base de la connaissance que possède le système perceptif sur ces corrélations. En remplaçant « contenir de l'information sur » par « information contenue dans », Gibson aurait subrepticement *réifié* le concept relationnel d'information. Il l'aurait traité « as a thing, rather than a relation » (p. 167). Une information *ne peut pas* affecter un système perceptuel. Seules des propriétés physiques le peuvent. Elles peuvent alors certes être « informatives sur quelque chose », mais seulement au moyen d'inférences. Car la *corrélation* elle-même qu'est l'information *ne peut pas* être un état d'un récepteur. Le problème est : « how (by what mental process) does the organism get from the detection of an informative property of the medium to the perception of a correlated property of the environment ? » Et la réponse est : par inférences. « X contient de l'information sur Y » est une relation *sémantique* et dépend donc de la façon dont X est mentalement représenté comme une prémisse d'inférences de X vers Y.

On voit que toute cette discussion (poussée beaucoup plus loin par les auteurs) repose sur le double préjugé que la réalité physique ne possède aucune propriété émergente et que ce qui est significatif doit nécessairement s'abstraire en sémantique et être produit par une intentionalité (la façon dont les représentations mentales dénotent). Par conséquent, il ne saurait exister dans l'environnement de structures intrinsèquement significatives encodables dans le signal lumineux.

L'existence d'une information morphologique géométriquement, *phéno*-physiquement et optiquement objective dément ce préjugé et

permet de fonder *un écologisme morphodynamique*. Gibson était dans le vrai avec son concept d'extraction d'invariants. Mais Fodor et Pylyshyn sont également dans le vrai en dénonçant chez lui un cercle vicieux. Il est effectivement vrai que « what we need is some criterion of being ecological *other than perceptibility* ». Mais ce critère, *c'est précisément le critère morphologique*. L'information morphologique *n'est pas sémantique*. Non relationnelle, elle est pourtant intrinsèquement significative. Elle peut affecter les systèmes sensoriels et perceptuels. A la suite de Thom, il faut méditer profondément sur ce statut *« sémio-physique »* des discontinuités qualitatives.

2. L'intentionalité

Un autre problème de base qu'une morphodynamique permet de résoudre sur le plan des principes est celui *de l'intentionalité* (cf. Petitot, 1984, 1986a, 1989b). On considère en général comme une évidence que l'intentionalité (la directionalité vers le monde externe) des représentations mentales est un fait sémantique. Selon nous, une telle approche, bien que traditionnelle, demeure irrémédiablement insuffisante. L'intentionalité est d'origine perceptive et les contenus sémantiques en héritent à travers la fondation de la structure conceptuelle dans le niveau 3D (au sens de Jackendoff-Marr). « Le problème des problèmes », comme dirait Husserl, est donc celui de l'intentionalité visuelle.

Or ce problème se trouve recevoir au niveau morphologique une réponse fort proche philosophiquement (mais évidemment fort éloignée mathématiquement) de celle qu'avait conçue Husserl. L'intentionalité visuelle se ramène essentiellement au passage des esquisses perceptives 2D à un objet identitaire 3D. Ce sont donc :

(i) le saut dimensionnel 2D $\rightarrow$ 3D ;

(ii) le principe de cohérence (le principe d'identité) que constitue l'objet pour la famille (l'espace fonctionnel) de ses esquisses,

qui en définissent le concept. Or nous avons vu que ce problème fondamental peut être désormais considéré comme résolu. L'intentionalité sémantique en perd du coup ses aspects aporétiques.

Cela montre bien toute l'importance de cette médiation morphologique entre le physique et le symbolique que nous avons tenté ici d'expliciter sur un exemple précis.

Jean PETITOT,
École des Hautes Études en Sciences Sociales.

BIBLIOGRAPHIE

AMIT (D.), 1989, *Modeling Brain Function,* Cambridge, Cambridge University Press.

ANDLER (D.), 1987, « Progrès en situation d'incertitude », *Le Débat, 47,* p. 5-25.

ARNOLD (V.), VARCHENKO (V.), GOUSSEIN-ZADE (S.), 1986, *Singularités des applications différentiables,* Moscou, Éditions Mir.

BALLARD (D. H.), BROWN (C. M.), 1982, *Computer Vision,* Englewood Cliffs, N. J., Prentice Hall.

BARROW (H. G.), TENENBAUM (J. M.), 1978, « Recovering Intrinsic Scene Characteristics from Image », in *Computer Vision Systems,* A. R. HANSON, E. M. RISEMAN eds, New York, Academic Press.

BIEDERMAN (I.), 1987, « Recognition-by-Components : A Theory of Human Image Understanding », *Psychological Review,* 94, *2,* p. 115-147.

BRADY (M.), 1982, « Computational Approaches to Image Understanding », *Computing Surveys,* 14, *1,* p. 3-71.

BRANDT (P.-A.), 1986, *La Charpente modale du Sens,* thèse de doctorat d'État, Université de Paris III.

BUSER (P.), IMBERT (M.), 1987, *Vision,* Paris, Hermann.

CHURCHLAND (P. M.), 1984, *Matter and Consciousness,* Cambridge, MA, MIT Press.

DESCLÉS (J.-P.), 1986, « Représentation des connaissances, archétypes cognitifs, schèmes conceptuels, schémas grammaticaux », *Actes sémiotiques,* VII, *69/70.*

FELDMAN (J. A.), 1985, « Four Frames Suffice : A Provisional Model of Vision and Space », *The Behavorial and Brain Sciences, 8,* p. 265-289.

FODOR (J. A.), PYLYSHYN (Z. W.), 1981, « How Direct Is Visual Perception ? Some Reflections on Gibson's " Ecological Approach " », *Cognition,* 9, p. 139-196.

FODOR (J. A.), 1984, *The Modularity of Mind,* Cambridge, MA, MIT Press.

GIBSON (J. J.), 1979, *The Ecological Approach to Visual Perception,* Boston, Houghton-Mifflin.

GRIMSON (W. E. L.), HILDRETH (E. C.), 1985, « Comments on Haralick 1984 », *IEEE, Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-7, p. 121-126.

HARALICK (R. M.), 1984, « Digital Step Edges from Zero Crossings of Second Directional Curvature », *IEEE, Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-6, p. 58-68.

HOFFMAN (D. D.), 1983, *Representing Shapes for Visual Recognition,* Doctoral Dissertation, MIT.

HOFFMAN (D. D.), RICHARDS (W. A.), 1984, « Parts of Recognition », *Cognition,* 18, p. 65-96.

HOFFMAN (D. D.), BENNETT (B. M.), 1986, « The Computation of Structure from Fixed-Axis Motion : Rigid Structures », *Biological Cybernetics,* 54, p. 71-83.

IKEUCHI (K.), « Shape from Regular Patterns », *Artificial Intelligence,* 22, p. 49-75.

JACKENDOFF (R.), 1983, *Semantics and Cognition,* Cambridge, MA, MIT Press.

JACKENDOFF (R.), 1987, *Consciousness and the Computational Mind,* Cambridge, MA, MIT Press.

KITCHER (P.), 1988, « Marr's Computational Theory of Vision », *Philosophy of Science,* 55, p. 1-24.

KOENDERINK (J.J.), VAN DOORN (A.J.), 1976, « The Singularities of the Visual Mapping », *Biological Cybernetics,* 25, p. 51-59.

KOENDERINK (J.J.), VAN DOORN (A.J.), 1979, « The Internal Representation of Solid Shape with Respect to Vision », *Biological Cybernetics,* 32, p. 211-216.

KOENDERINK (J.J.), VAN DOORN (A.J.), 1986, « Dynamic Shape », *Biological Cybernetics,* 53, p. 383-396.

KOENDERINK (J.J.), VAN DOORN (A.J.), 1987, « Representation of Local Geometry in the Visual System », *Biological Cybernetics,* 55, p. 367-375.

KOSSLYN (S. M.), 1980, *Image and Mind,* Cambridge, MA, Harvard University Press.

LANGACKER (R.), 1987, *Foundations of Cognitive Grammar,* Stanford University Press.

LE DÉBAT, 1987, « Une nouvelle science de l'esprit », *Le Débat, 47.*

LTC, 1989, *Logos et Théorie des catastrophes,* Colloque de Cerisy à partir de l'œuvre de René Thom, Jean PETITOT, éd., Genève, Éditions Patiño.

MARR (D.), 1982, *Vision,* San Francisco, Freeman.

MEYER (Y.), 1989, « Ondelettes, filtres miroirs en quadrature et traitement numérique de l'image », *Gazette des mathématiciens,* 40, p. 31-42.

MINGOLLA (E.), TODD (J.T.), 1986, « Perception of Solid Shape from Shading », *Biological Cybernetics,* 53, *3,* p. 137-151.

OUELLET (P.), 1987, « Une physique du sens », *Critique, 481/482,* p. 577-597.

P.D.P., 1986, *Parallel Distributed Processing,* David E. RUMELHART, James L. MCCLELLAND eds, Cambridge, MIT Press.

PETITOT (J.), 1977, « Topologie du carré sémiotique », *Études littéraires,* p. 347-428, Québec, Université de Laval.

PETITOT (J.), 1979, « Hypothèse localiste et Théorie des catastrophes », in *Théories du langage, théories de l'apprentissage,* M. PIATTELLI, éd., Paris, Le Seuil.

PETITOT (J.), 1982, *Pour un schématisme de la structure,* thèse de doctorat d'État, Paris, E.H.E.S.S.

PETITOT (J.), 1983, « Théorie des catastrophes et structures sémio-narratives », *Actes sémiotiques,* V, *47/48,* p. 5-37.

PETITOT (J.), 1984, « La lacune du contour », *Anàlise,* 1, *1,* p. 101-140, Lisbonne.

PETITOT (J.), 1985, *Morphogenèse du Sens,* Paris, Presses Universitaires de France.

PETITOT (J.), 1986 a, « Le " morphological turn " de la phénoménologie », *Document du C.A.M.S.,* Paris, E.H.E.S.S.

PETITOT (J.), 1986b, « Épistémologie des phénomènes critiques », *Document du C.A.M.S.*, Paris, E.H.E.S.S.

PETITOT (J.), 1988, « Approche morphodynamique de la formule canonique du mythe », *L'Homme,* 106-107, XVIII *(2-3),* p. 24-50.

PETITOT (J.), 1989a, « Éléments de dynamique modale », *Poetica et Analytica,* 6, p. 44-79, Université d'Aarhus.

PETITOT (J.), 1989b, « Structuralisme et Phénoménologie », *LTC,* 1989, p. 345-376.

PETITOT (J.), 1989c, « On the Linguistic Import of Catastrophe Theory », *Semiotica,* 74, *3/4,* p. 179-209.

PETITOT (J.), 1989d, « Catastrophe Theory and Semio-Narrative Structures », in *Paris School of Semiotics,* P. PERRON, F. COLLINS, eds, Amsterdam, John Benjamins, p. 177-212.

PETITOT (J.), 1989e, « Morphodynamics and the Categorical Perception of Phonological Units », *Theoretical Linguistics,* 15, *1/2,* p. 25-71.

PETITOT (J.), 1989f, « Hypothèse localiste, Modèles morphodynamiques et théories cognitives : remarques sur une note de 1975 », *Semiotica,* 77, *1/3,* p. 65-119.

PETITOT (J.), 1989g, « Forme », *Encyclopaedia Universalis,* XI, p. 712-728, Paris.

PETITOT (J.), 1989h, « La modélisation : formalisation ou mathématisation ? L'exemple de l'approche morphodynamique dans les sciences du langage », in *Perspectives méthodologiques et épistémologiques dans les sciences du langage,* M.J. REICHLER-BÉGULIN, éd., Bern, Peter Lang, p. 205-220.

PETITOT (J.), 1989i, « Why Connectionism Is Such a Good Thing ? », in *Workshop Connectionism and Language,* San Marino, Università degli Studi.

PINKER (S.), 1984, « Visual Cognition : An Introduction », *Cognition,* 18, p. 1-63.

PINKER (S.), ed., 1984, *Visual Cognition, Cognition,* 18, Cambridge, MA, MIT Press.

POGGIO (T.), 1984, « Vision by Man and Machine », *Scientific American,* 250, *4,* p. 68-78.

PRÉFACES, 1988, « Un tournant cognitif dans les sciences humaines », *Préfaces, 10,* p. 67-105.

PROUST (J.), 1987, « L'intelligence artificielle comme philosophie », *Le Débat,* 47, p. 88-102.

PYLYSHYN (Z.), 1986, *Computation and Cognition,* Cambridge, MA, MIT Press.

RICHTER (J.), ULLMAN (S.), 1986, « Non-Linearities in Cortical Simple Cells and the Possible Detection of Zero Crossings », *Biological Cybernetics,* 53, *3,* p. 195-202.

SHEPARD (R.N.), COOPER (L.A.), 1982, *Mental Images and their Transformations,* Cambridge, MA, MIT Press.

SMOLENSKY (P.), 1988, « On the Proper Treatment of Connectionism », *The Behavioral and Brain Sciences,* 11, p. 1-74.

STILLINGS (N.A.), *et al.,* 1987, *Cognitive Science. An Introduction,* Cambridge, MA, MIT Press.

TALMY (L.), 1978, « Relation of Grammar to Cognition », in *Proceedings of TINLAP-2,* D. WALTZ, ed., Urbana, University of Illinois.

TALMY (L.), 1983, « How Language Structures Space », in *Spatial Orientation : Theory, Research and Application*, H. PICK, L. ACREDOLO, eds, New York, Plenum Press.

TALMY (L.), 1985, « Force Dynamics in Language and Thought », *Parasession on Causatives and Agentivity*, Chicago Linguistic Society, 21st. Regional Meeting.

THOM (R.), 1972, *Stabilité structurelle et Morphogenèse*, New York, Benjamin, Paris, Édiscience.

THOM (R.), 1978, « Morphogenèse et Imaginaire », *Circé, 8-9*, Paris, Éditions Lettres Modernes.

THOM (R.), 1980, *Modèles mathématiques de la Morphogenèse*, 2e éd., Paris, Christian Bourgois.

THOM (R.), 1988, *Esquisse d'une Sémiophysique*, Paris, Inter-Éditions.

ULLMAN (S.), 1979, *The Interpretation of Visual Motion*, Cambridge, MA, MIT Press.

ULLMAN (S.), 1984, « Visual routines », *Cognition*, 18, p. 97-159.

WILDGEN (W.), 1982, *Catastrophe Theoretic Semantics*, Amsterdam, Benjamins.

ZEEMAN (Ch.), 1977, *Catastrophe Theory*, Massachusetts, Addison-Wesley.