Chapitre 3 :
Ancrage des concepts

Introduction

Les concepts ne sont pas des entités suspendues dans un éther cognitif. En tant que supports mentaux de la signification, ils constituent un intermédiaire entre un mot et l'objet ou la situation que ce mot permet de désigner dans le monde perçu. De ce fait, les concepts doivent être ancrés dans nos expériences. En d'autres termes, le système conceptuel, qui a en charge la construction du sens et héberge les significations, doit posséder une interface avec la perception pour que les significations puissent porter sur les données de nos expériences.

3.1. Les données du problème

À partir du moment où les concepts sont considérés comme des représentations mentales, le problème de leur ancrage se pose. Or, la manière dont cet ancrage peut être réalisé fait l'objet de débats opposant des positions bien distinctes. Ces débats touchent à des questions fondamentales comme la nature du contenu et de la forme des représentations conceptuelles, et sont indissociables du type de modélisation auquel on souscrit. Dans ce chapitre, notre objectif n'est pas d'argumenter et de prendre parti, mais plutôt de présenter certaines positions présentes dans la littérature concernant cette question centrale de l'ancrage. Nous ne chercherons pas à être exhaustive, ne retenant ici que les théories qui interfèrent avec notre problématique. Nous commençons, dans cette section, par clarifier certains points de vocabulaire.

Représentations, modélisation et ancrage

La notion de représentation est avant tout un outil théorique pour l'explication du fonctionnement cognitif. La variété des comportements humains ne saurait être expliquée par une simple correspondance avec des configurations de stimuli. La majorité des modèles de la cognition humaine utilisent la notion de représentation pour médiatiser le lien complexe entre l'expérience du sujet et le comportement observable qu'il produit. Toutefois, ces modèles se divisent en deux groupes, selon le type de calcul qu'ils postulent, si bien que la notion de représentation est utilisée dans deux sens radicalement différents.

> Standard explanations of how systems come to exhibit sophisticated cognitive performances advert to internal representations. Computationalists take representations to be static configurations of symbol tokens. Dynamicists conceive representations very differently. They find their representations among the kind of entities that figure in [dynamic system theory] [...] Unlike digital computers, dynamical systems are not inherently representational. A small but influential contingent of dynamicists have found the notion of representation to be dispensable or even a hindrance for their particular purposes. (□AN GELDER 1998 [105] p. 622)

Un modèle computationnel de la cognition repose sur l'application séquentielle de règles explicites à des représentations symboliques. Les représentations jouent le rôle de *tokens*, de simples "jetons" manipulés par le système d'après leur forme. Les règles, dans la mesure où elles sont explicites, sont elles-mêmes constituées de *tokens*. Ces *tokens* sont supposés avoir une existence physique identifiable dans le système (NEWELL 1980 [78]). Cette nécessité de considérer l'implantation matérielle des *tokens* mise à part, un modèle

symbolique constitue une description qui se situe à un niveau relativement abstrait, celui des manipulations symboliques formelles telles que celles qu'effectue une machine de Turing. Cependant, du fait de l'existence matérielle des symboles manipulés, un tel modèle est bien davantage qu'un simple outil théorique permettant de prédire l'évolution du système. Il existe un isomorphisme supposé entre les opérations décrites dans le modèle et les mécanismes physiques qui se déroulent dans le système. Dans un tel schéma, la notion de représentation est donc incontournable. Non seulement les *tokens* constituent les ingrédients fondamentaux du modèle, mais le modèle prédit même la possibilité de caractériser ces *tokens* de manière indépendante, par exemple par une technique de neurophysiologie.

Cette manière de concevoir le traitement cognitif introduit automatiquement une distinction de type syntaxe *versus* sémantique. La manipulation des représentations est une mécanique formelle, une syntaxe, qui dépend exclusivement de la structure de ces représentations. Pour que cette mécanique ne tourne pas à vide, il faut bien que ces représentations soient ancrées, soit dans les mécanismes perceptifs, soit dans les mécanismes effecteurs. Elles possèdent, de ce fait, une sémantique : elles sont interprétables dans le domaine des perceptions ou dans celui des actions.

> We need the syntactic or the symbolic level because we must preserve certain interpretations over mental operations […] This we can do only if we have a semantic function whose definition has access to the generative structure of the symbolic expressions. […] To count as a computation rather than simply any functionally described physical system, [a machine] must contain symbols that are interpreted. […] This quality of symbols and of computational states, whereby they can consistently be given a semantic interpretation, is not the only thing that makes useful computation possible; but it is one of the most important characteristics shared by computation and cognition.
>
> (PYLYSHYN 1984 [88] p. 62)

Les propriétés sémantiques doivent être compatibles avec la syntaxe : à partir d'un premier ensemble de représentations directement ancrées, la mécanique syntaxique engendre de nouvelles représentations qui peuvent être, à leur tour, sémantiquement interprétées. Cette contrainte impose que les représentations intermédiaires, celles qui interviennent dans la chaîne de traitement symbolique, soient également interprétables, même si elles ne sont pas directement associées, par le biais de l'expérience, à la perception ou aux actions[1] (HARNAD 1990 [46]). Comme elles possèdent une structure formelle interne, elles peuvent hériter leur ancrage de leurs constituants et du calcul qui les a produites.

> What is the representation of a zebra? [In our example] it is just the symbol string "horse & stripes". But because "horse" and "stripes" are grounded in their respective iconic and categorical representations, "zebra" inherits the grounding, through its grounded *symbolic* representation. In principle, someone who had never seen a zebra (but had seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone (plus the nonsymbolic representations of horses and stripes that ground it). Once one has the grounded set of elementary symbols provided by a taxonomy of names (and the iconic and categorical representations that give content to the names and allow them to pick out the objects they identify) the rest of the symbol strings of a natural language can be generated by symbol composition alone, and they will all inherit the intrinsic grounded of the elementary set. (HARNAD 1990 [46] p. 343)

[1] Cela n'empêche pas l'introduction de tokens non interprétables en tant que tels. Par exemple, dans un système logique classique, les symboles de prédicats ou les constantes reçoivent une interprétation directe, au contraire des symboles logiques dont la fonction se limite à diriger l'interprétation des expressions complexes.

La situation est tout autre si l'on adopte une théorie dynamique de la cognition. Les théories dynamiques considèrent le modèle formel comme un outil de prédiction qui simule le système physique, sans postuler l'existence d'un isomorphisme. Ainsi, les équations de Kepler ne sont pas censées posséder la moindre contrepartie matérielle dans le système des planètes. La description formelle à l'aide de symboles désignant la position d'une planète, sa vitesse ou l'aire du rayon vecteur est propre au modèle, et ne s'applique que métaphoriquement au système matériel. La notion de représentation, contrairement aux modèles symboliques, est ici facultative. Une partie des défenseurs des systèmes dynamiques sont même anti-représentationnalistes, considérant que rien, dans le système, n'est interprétable dans les termes d'un domaine extérieur au système.

> L'idée fondamentale est donc que les facultés cognitives sont inextricablement liées à l'historique de ce qui est vécu, de la même manière qu'un sentier au préalable inexistant apparaît en marchant. L'image de la cognition qui s'ensuit n'est pas la résolution de problèmes au moyen de représentations, mais plutôt le faire–émerger créateur d'un monde, avec la seule condition d'être opérationnel : elle doit assurer la pérennité du système en jeu.
>
> (□ARELA 1988 [106] p. 111)

Cependant, une telle position ne représente pas un consensus parmi les défenseurs de l'approche dynamique. Notamment, certains modèles connexionnistes s'en éloignent en introduisant la notion de représentation distribuée.

> While [behaviorist accounts of behavior] do involve simple mechanisms of learning, there is a crucial difference between our models and the radical behaviorism [...]. In our models, we are explicitly concerned with the problem of internal representation and mental processing, whereas the radical behaviorist explicitly denies the scientific utility and even the validity of the consideration of these constructs. [Our connectionist models] all concern internal mechanisms for activating and acquiring the ability to activate appropriate internal representations. (RUMELHART & MCCLELLAND 1986 [93] p. 121)

Certaines caractéristiques des modèles dynamiques méritent l'étiquette de représentation. On peut appeler représentation l'état global du système, celui qui résume à lui seul tout son passé. Ainsi, dans un système de planètes considéré du point de vue de la mécanique newtonienne, la position du centre de gravité des planètes et leur vitesse suffit à résumer l'état du système et à permettre la prédiction de ses états futurs. Dans un modèle de réseau de neurones, il en est de même de l'ensemble de valeurs constitué par les poids synaptiques et les activités des neurones. Il y a dans ce cas une seule représentation globale, dont les sous-ensembles peuvent être également considérés comme des représentations. Ainsi, les données relatives à une planète constituent la représentation associée à cette planète. De même, on peut appeler représentation l'état d'activité d'un sous-ensemble de neurones.

> Should we identify one's conceptual framework with the configuration of synaptic *weights* in one's brain? Or with the *partitions* they effect across the activation vector space of the assembled neurons to which they connect? Or perhaps with the overall input–output *function* that the network comes to instantiate? The weights uniquely dictate both the partitions and the function, but despite the functional primacy of the weights, there are good reasons for identifying the partitions, and the function they serve, as reflecting most directly the antecedent notion of a "conceptual framework". [...] At least for now, therefore, let us adopt the partitions and the functions they serve as the closest available neural analogue of what the philosophical tradition conceives as our "conceptual framework".
>
> (CHURCHLAND 1989 [19] p. 232)

Dans un réseau de neurones sujet à des phénomènes de synchronisation, le choix du support matériel des représentations laisse moins de doute. Tout ensemble synchrone de neurones peut être considéré, de manière fort naturelle comme une représentation, et ceci d'autant plus facilement que la synchronisation est parfois invoquée comme corrélat de la conscience (CRICK & KOCH 1990 [23]).

Les systèmes dynamiques comme les réseaux de neurones présentent la propriété de posséder des attracteurs. Ainsi, un réseau de neurones placé dans certaines conditions initiales, va évoluer spontanément vers un état d'équilibre. Dans la mesure où ils sont façonnables par apprentissage, ces attracteurs sont interprétables par un observateur : ils sont ancrés dans la configuration d'entrée du système qui a permis leur stabilisation par apprentissage (HOPFIELD 1982 [48]). Il est donc intéressant de considérer que les attracteurs d'un tel système dynamique constituent autant de représentations.

Ce type d'ancrage, propre aux systèmes dynamiques, a cependant ses limites. C'est ainsi que pour certains auteurs, la sémantique des attracteurs a peu de choses à voir avec les entrées du système. Le rôle de ces entrées est ramené à celui de simples perturbations. Le système possède un certain nombre d'attracteurs, mais ces attracteurs ne dépendent que de manière fortuite de l'histoire des interactions du système avec l'extérieur. D'où l'idée selon laquelle le système crée des significations pour lui-même (□ARELA 1988 [106]). Certains attracteurs peuvent sembler posséder une signification pour la raison qu'ils sont corrélés à des configurations de stimuli, mais ce genre de signification n'existerait que dans l'œil du modélisateur (SKARDA & FREEMAN 1987 [95]). Selon cette manière de concevoir les systèmes cognitifs, il ne saurait donc y avoir d'ancrage proprement dit.

Lorsque l'on s'intéresse au langage, il est difficile de souscrire d'emblée à la position qui vient d'être invoquée. Il semble que les mots perçus dans le message de l'interlocuteur aient un effet moins aléatoire que celui de simples perturbations. Puisque l'interaction langagière est possible, il est permis de supposer qu'elle ne résulte pas de la confrontation entre deux systèmes "autistes" ne connaissant que leurs propres significations internes. Nous considérons donc, dans ce document, que l'hypothèse représentationnaliste mérite d'être explorée, que ce soit sous sa forme computationnelle ou dans sa réalisation dans les systèmes dynamiques. Nous nous intéresserons donc à des systèmes dont certains états peuvent être qualifiés de représentationnels. Nous nous demanderons comment ces états peuvent entretenir un rapport d'ancrage avec d'autres états du système ou avec certaines configurations des entrées et des sorties. Nous constaterons, dans ce chapitre, que cette question de l'ancrage est particulièrement problématique.

Représentations : le contenu

La notion de représentation constitue un outil commode, peut-être incontournable, pour élaborer et exprimer un modèle de la sémantique du langage. Cependant, l'enjeu dépasse la question de la modélisation du langage. La question de la nature des représentations est au centre d'un débat philosophique concernant leur propriété intentionnelle, propriété attribuée à certains phénomènes cognitifs. Certains états mentaux semblent posséder la propriété remarquable d'être obligatoirement à propos de quelque chose. Ainsi, une grande partie des états conscients, par exemple l'état correspondant au fait de voir orange ou d'avoir mal, sont à propos de quelque chose : ils ont un contenu. Ainsi, l'état conscient associé à la couleur orange sera lié à la vision d'un livre ou de la surface d'un mur ; une douleur donnée sera ressentie comme si elle se trouvait dans le pied ou dans une dent. D'autres types d'états mentaux possèdent manifestement cette propriété d'avoir un contenu. Par exemple, une croyance ou un désir n'existe cognitivement, la plupart du temps, que si elles sont à propos d'un état de choses, par exemple le fait de réussir un examen. La question de l'intentionnalité

se pose donc pour plusieurs types d'états mentaux, parmi lesquels les représentations mentales associées aux expressions langagières occupent une place de premier plan. Ainsi, la représentation mentale associée au mot arbre semble être à propos de quelque chose : l'image d'un arbre qui se trouve dans la cour ou cet arbre lui-même. Dans le cas des représentations mentales impliquées dans le traitement sémantique du langage, la propriété d'intentionnalité, le fait de posséder un contenu, permet d'expliquer la possibilité même de la communication : si les expressions langagières renvoient à des représentations mentales qui ont un contenu, les expressions langagières servent à communiquer à propos de ce contenu. Comme nous allons le constater, les choses ne sont pas si simples.

Il est important de noter que la notion technique de représentation n'est pas obligatoirement liée à celle de contenu. Il est par exemple courant, dans les modèles de la syntaxe, de postuler l'existence de représentations dont la raison d'être est interne au système et qui n'ont pas de contenu. C'est ainsi que le "pronom" PRO est invoqué pour expliquer certains phénomènes syntaxiques, notamment l'existence des impersonnels. Dans la phrase il pleut, l'impersonnel il donne une réalité phonologique au pronom PRO, ce qui n'est pas le cas dans les langues dites *pro-drop* comme le persan. Le pronom PRO est une représentation, un *token* qui entre dans le calcul formel proposé par le modèle et dont la présence semble incontournable pour expliquer certains phénomènes syntaxiques. En tant que représentation, le pronom PRO est supposé avoir une réalité cognitive. Pourtant, rien ne permet, dans le modèle, de lui allouer un contenu.

À l'autre extrême, on pourrait penser que certaines représentations ont une intentionnalité tellement affirmée que leur contenu, ce sur quoi elles portent, se retrouve dans leur forme. Selon cette idée, la représentation serait une re-présentation, une sorte de copie d'une autre chose qui lui préexisterait. Cette idée n'est pas entièrement sans fondement. On sait par exemple que certaines propriétés topologiques sont conservées entre la rétine, les relais thalamiques et les différentes cartes corticales du système visuel, si bien que l'image d'une pomme qui se forme sur la rétine est, par certains aspects, copiée dans certaines aires du cortex. Si l'on conçoit le fonctionnement cognitif comme un traitement de type purement analogique, ces images corticales peuvent constituer d'authentiques re-présentations qui peuvent déclencher d'autres traitements, par exemple une réponse motrice consistant à cueillir la pomme. Cependant, de nombreux autres types de traitements, notamment les traitements de type computationnel, supposent des niveaux de représentation dans lesquels les différentes parties des images corticales sont intégrées et ne sont pas conservées. Dans ce cas, la représentation n'entretient plus aucun rapport de ressemblance avec l'image rétinienne. Dans la plupart des modèles, l'intentionnalité des représentations ne suppose pas que la forme de la représentation re-présente son contenu. On doit plutôt concevoir le lien intentionnel comme un renvoi : la représentation renvoie à son contenu, il est possible d'obtenir l'un à partir de l'autre, sans pour autant que les deux se ressemblent. Dans ce cas, l'intentionnalité rejoint d'emblée la notion d'ancrage. Il est important d'insister sur le fait qu'il est parfaitement possible de postuler l'existence, dans le système cognitif, de représentations dotées d'une intentionnalité sans que celles-ci soient des re-présentations.

La question que nous essayons de traiter dans ce travail concerne la nature des représentations mentales associées aux expressions langagières. S'agit-il de simples pions manipulés selon une mécanique aveugle ? Dans ce cas, le caractère intentionnel d'une expression langagière pourrait n'être qu'une illusion. Il serait produit *post hoc* par une opération d'interprétation. La phrase donne-moi le crayon qui est sur la table exprime un souhait à propos de la possession du crayon en question. Ce souhait pourrait n'être qu'une interprétation produite par l'interlocuteur et n'avoir aucune réalité cognitive interne à l'esprit du locuteur, pas plus que le souhait exprimé par un ordinateur qui affiche la phrase j'ai faim sur son écran. Le caractère intentionnel d'une expression langagière serait ainsi dérivé de l'acte

interprétatif de l'observateur et n'exprimerait aucune propriété intrinsèque au processus cognitif qui a produit cette expression (DAVIDSON 1984 [25]). Sur de telles bases, une théorie du langage se contentera de chercher le caractère intentionnel de la communication langagière dans l'acte de traduction dans une autre langue, dans la prédiction des comportements ou dans l'attribution des valeurs de vérités aux énoncés par l'auditeur.

Ce type d'approche peut convenir pour une théorie behavioriste de la communication langagière. Malheureusement, le pouvoir prédictif que l'on peut attendre d'une telle théorie est limité, si on le compare à celui des théories représentationnalistes. Lorsqu'un locuteur énonce la phrase donne-moi le crayon qui est sur la table, sa phrase ne saurait être le produit d'un simple processus associatif. Il est plausible de supposer qu'elle résulte d'un processus impliquant des représentations. Par exemple, le locuteur désire écrire une lettre, se retrouve dans l'embarras de ne pas avoir de crayon, aperçoit le crayon qui est sur la table, et forme le plan qui consiste à demander à l'interlocuteur un petit service. Les éléments de ce processus, tel qu'on peut scientifiquement le reconstituer, se situent au même niveau que la description qui vient d'être esquissée. Il est, de plus, plausible de postuler qu'il s'agit de représentations qui renvoient aux objets perçus comme lettre, crayon, table. Les règles qui dirigent ce calcul peuvent s'exprimer au même niveau. Par exemple, une telle règle stipule de résoudre toute incompatibilité détectée entre attitudes propositionnelles. Ainsi, le locuteur résout une incompatibilité entre le désir d'écrire la lettre et le constat qu'il n'en a pas les moyens matériels. Si l'interlocuteur réagit en prononçant la phrase ce n'est pas un crayon, mais un porte-mine, c'est qu'il conçoit une incompatibilité entre ce qu'on lui dit et ce qu'il croit pour l'avoir perçu. De même, l'interlocuteur qui répond par la phrase il vaut mieux écrire ta lettre au stylo résout l'incompatibilité entre la croyance que la lettre va être écrite au crayon et le désir de conférer à la lettre un caractère plus officiel. De tels calculs opèrent sur des représentations dotées d'une propriété intentionnelle. Renoncer à l'existence cognitive de telles représentations conduit à renoncer au pouvoir prédictif de ce type de calcul.

À partir du moment où l'on recherche un modèle de la génération et de la compréhension des expressions langagières qui met en jeu plusieurs niveaux de représentations, on peut s'interroger sur l'intentionnalité propre à chaque niveau. Comme nous venons de le voir, le niveau où les phrases sont interprétées suppose une forme d'intentionnalité, celle des attitudes propositionnelles. On pourrait envisager d'en rester là, en supposant que toutes les représentations qui se situent en-dessous, les représentations associées aux syntagmes, aux mots, aux syllabes, *et cætera*, sont dépourvues d'intentionnalité. Dans ce texte, nous allons considérer en détail l'hypothèse selon laquelle il existe des représentations mentales associées aux entités lexicales, les concepts, possédant un caractère intentionnel propre, qui peut être qualifié de sémantique[2]. Les entités sub-lexicales, comme la syllabe, peuvent recevoir d'autres types de contenu, par exemple une forme acoustique. Le point de vue que nous adopterons dans la majeure partie de ce texte est que les concepts sont les représentations mentales minimales dotées d'une sémantique, c'est-à-dire qu'ils s'expriment dans le monde perçu.

> On all hands, then, concepts serve both as the domains over which the most elementary mental processes are defined, and as the most primitive bearers of semantic properties. Hence their centrality in representational theories of mind. (FODOR 1994 [36] p. 96)

Une position fréquemment adoptée consiste à considérer que le contenu des concepts se situe dans le "monde", considéré comme objectivement accessible. Cette position semble raisonnable, dans la mesure où l'on utilise les concepts pour communiquer à propos des

[2] Nous reviendrons sur ce point en introduisant la notion plus stricte de concept lexical (CF. CHAPITRE 6).

entités et des situations du monde. Les concepts, en tant que représentations mentales, auraient ainsi cette propriété de renvoyer à des entités indépendantes de l'esprit humain. Le postulat de l'existence de ce type d'entité est qualifié de réaliste. Les candidats pour ce type d'entité sont des objets concrets, des objets abstraits, des classes d'objets, des attributs d'objets, des propriétés, des faits, des états de faits, *et cætera*. La question de l'existence indépendante de ces entités et de leurs qualités est une question métaphysique qui divise les penseurs en deux groupes, selon qu'ils sont réalistes ou non par rapport aux catégories universelles dans le monde. La contrepartie sémantique de la question revient à se demander si l'on établit un lien entre les concepts et de telles entités externes, ce qui constitue le point de vue externaliste, ou si le contenu des concepts est lui-même une entité mentale, ce qui constitue le point de vue internaliste.

Nous n'entrerons pas dans le débat métaphysique du réalisme, car son issue n'a pas de conséquence sur notre thèse. Nous choisissons, dans ce texte, d'éviter l'hypothèse de l'existence d'un monde structuré indépendant de l'esprit. En revanche, nous adopterons une position, empruntée à la tradition kantienne, qualifiée de réalisme empirique. Le réalisme empirique affirme l'existence des entités mentales qui consistent en notre expérience. Nous parlons ici des expériences au sens large, ce qui inclut la perception mais aussi les sensations et les attitudes. En ce qui concerne le monde extérieur, la seule conviction que nous pouvons avoir quant à son existence est liée au sentiment intuitif que quelque chose nous résiste. L'intuition d'une résistance qui contrarie nos souhaits ou notre volonté nous conduit à faire l'hypothèse d'une réalité extérieure, sans que nous ne puissions dire quoi que ce soit sur la cause précise de cette résistance. Les seules informations auxquelles nous avons accès, en tant que sujet connaissant, nous sont fournies par l'état d'activité de nos capteurs sensoriels. C'est à partir de ces informations et à propos de ces informations que nous formons des expressions langagières.

En marge de cette question métaphysique se pose le problème de choisir entre les versions internaliste et externaliste de l'explication de l'intentionnalité des concepts. Les concepts, en tant qu'entités mentales intentionnelles, portent-ils sur d'autres entités mentales ou portent-ils sur des entités qui sont dans le monde ? Il ne s'agit pas de répondre à cette question dans l'absolu. Lorsque l'on aborde cette question sous l'angle de la modélisation, la tâche est d'expliquer l'établissement du lien intentionnel pour un sujet qui est le produit exclusif d'une phylogenèse et d'une ontogenèse, en renonçant à toute autre source d'information. Lorsque l'on parle de l'objet "ballon" ou de la propriété "rouge" ou de la situation "jeu", il faut lier les concepts correspondants, *id est* les représentations mentales associées à ces mots, à des phénomènes vécus par le sujet. Là encore, l'hypothèse d'un monde structuré peut être évitée, puisque le sujet doit l'intentionnalité de ses représentations mentales à ses dispositions biologiques et aux seules données contenues dans ses expériences[3]. Nous sommes donc contrainte de renoncer, pour des raisons épistémologiques, aux facilités que procure l'externalisme pour adopter une position que nous appelons internalisme méthodologique : les concepts acquièrent leur contenu dans le domaine des entités mentales associées aux expériences du sujet[4]. Cette position ne présuppose aucun jugement concernant la correspondance entre les concepts et le monde. Elle consiste à proposer d'expliquer la mise en place des concepts par le seul fait des dispositions mentales et des expériences du sujet. C'est dans cette perspective que la notion d'intentionnalité rejoint celle d'ancrage. Le caractère intentionnel des concepts résulte ainsi du fait qu'ils sont, en tant que représentations mentales, ancrés dans les données de l'expérience.

[3] Nous discutons plus loin la possibilité d'une intentionnalité innée qui serait le résultat de la phylogenèse, pour conclure que cette possibilité ne change pas fondamentalement le point de vue que nous adoptons ici.

[4] Nous démarquons plus loin notre position de celle qui est habituellement appelée internaliste.

Dans la perspective de l'internalisme méthodologique, l'intentionnalité des concepts se présente comme un phénomène d'évocation. Les concepts sont intentionnels car il existe des expériences qui les évoquent et qu'ils évoquent. La question de l'intentionnalité des concepts se ramène donc, du point de vue de la modélisation, au problème d'expliquer l'interface entre le système conceptuel et d'autres systèmes de la cognition, notamment la perception, qui produisent les entités mentales associées aux expériences. Cette manière de procéder peut sembler en contradiction avec la position externaliste. C'est en partie le cas. Ceux qui adoptent le point de vue réaliste considéreront qu'il existe trois domaines : les choses du monde, les expériences et les concepts. Ils peuvent imaginer deux interfaces, entre le monde et les expériences d'une part, entre les expériences et le système conceptuel d'autre part. Le projet externaliste consiste notamment à établir un lien intentionnel objectif entre les concepts et le monde en composant ces deux interfaces. Ainsi, là où l'internalisme méthodologique explore des mécanismes d'évocation des concepts dans les données de l'expérience, le point de vue externaliste propose des conditions d'application des concepts aux entités du monde extérieur. Cependant, la définition de la première interface, entre le monde et les expériences, suppose que l'on dispose d'un "oracle", un regard externe qui embrasse les deux domaines. Dans une perspective matérialiste, cet oracle est habituellement constitué par la science. Ceci signifie que les catégories attribuées au monde sont fixées par nos connaissances scientifiques actuelles ou potentielles. Il ne reste ensuite qu'à établir un lien entre ces catégories "réifiées" et la cognition. Une telle méthodologie est toutefois problématique, si l'on admet que la science n'est qu'un produit de la cognition humaine. Il y a donc pétition de principe, puisque c'est l'analyse des capacités cognitives qui doit nous amener aux lois qui gouvernent l'apparition de ces catégories du savoir. On ne saurait ancrer ce qui constitue une partie fondamentale de la cognition humaine, le système conceptuel, dans les catégories provisoires produites par l'activité de la communauté scientifique. Lorsque l'on cherche à élaborer un modèle plausible de la compétence sémantique des humains, on ne peut pas accepter que les nouvelles théories que des physiciens des particules peuvent produire dans leur laboratoire puissent avoir une influence quelconque sur cette compétence. Toutefois, les théories externalistes nous intéresseront dans la mesure où, en dédoublant l'interface concept - monde, nous pouvons leur donner une interprétation internaliste, en particulier en ce qui concerne l'acquisition des concepts. Là où une théorie externaliste postule l'existence d'entités indépendantes de la cognition pour établir un lien intentionnel avec les concepts nouvellement acquis, nous ne nous intéresserons qu'au seul lien d'ancrage entre les concepts et les expériences que ces entités indépendantes sont supposées susciter.

On pourrait reprocher à l'internalisme méthodologique de supprimer la notion de référence. Dans un cadre externaliste, un mot du lexique renvoie à une entité du monde qui constitue la référence de ce mot. L'existence des références externes pour les mots est une hypothèse forte, motivée entre autres par le souci d'expliquer que la communication au moyen du langage soit possible. L'efficacité du langage, dans ce cadre, s'explique de la manière la plus simple qui soit : en prononçant un mot, le locuteur désigne une entité du monde à son interlocuteur. Ainsi, le projet externaliste peut se focaliser, d'une part, sur l'établissement du lien référentiel des mots aux entités du monde, et d'autre part, sur la propagation de ce lien à toutes les expressions langagières. En supprimant toute référence directe au monde, l'internalisme méthodologique semble renoncer à expliquer la communication. Si les mots n'évoquent que des représentations mentales, si l'on ne peut pas se référer à des entités du monde, il semble impossible d'établir une corrélation entre les états mentaux des interlocuteurs. Nous considérons cependant que cette conclusion est erronée et que le problème de la communication langagière mérite précisément une analyse d'un point de vue strictement cognitif, sans un recours aux postulats métaphysiques sur l'existence des entités indépendantes de l'esprit humain. Nous tâcherons plus tard de développer notre point

de vue sur ce sujet (CF. CHAPITRE 9). Dans le cadre que nous adoptons, la notion de référence externe est remplacée par une référence interne. Par exemple, la référence du mot pomme pourra être, dans certains contextes, une image ou un goût, plutôt qu'un élément de nomenclature botanique comme l'exigerait sans doute une position externaliste matérialiste.

La question du contenu d'un concept porte donc sur la référence du terme auquel est associé ce concept. Cette référence appartient à un autre domaine que celui des concepts, le domaine général des expériences. Or, le fait que le concept soit ainsi ancré ne nous renseigne pas sur sa propre nature. Après avoir considéré la cible de l'ancrage, nous abordons maintenant la question d'analyser la forme de ce qui est ancré.

Représentations : la forme

Les modèles représentationnalistes postulent l'existence de représentations de manière à expliquer et à prédire le comportement du système cognitif. Dans la mesure où l'on cherche à décrire ce comportement à l'aide de mécanismes généraux, il est indispensable que les représentations soient dotées d'une structure interne. C'est le moyen par lequel ces mécanismes peuvent les discerner. Lorsque des mécanismes généraux agissent sur des représentations, ils le font d'après leur forme, *id est* d'après leur structure interne, ce qui permet de les différencier des autres représentations. Dans un modèle dynamique, la structure de la représentation résume une partie de l'état du système et est nécessaire, au sein du modèle, pour prédire son trajet avenir. Dans un système computationnel, la notion de forme est encore plus importante, puisqu'elle est supposée avoir une contrepartie matérielle. Les règles du modèle, qui ne sont pas limitées à un simple moyen de description mais sont censées être constitutives du système, manipulent des représentations, elles-aussi matériellement constitutives du système, d'après leur forme. On peut, dans ce cas, considérer que la forme des représentations résume leur propriété causale au sein du système.

Il est difficile de concevoir une théorie de la communication langagière qui soit non représentationnaliste, tant on est frappé par l'apparente complexité des processus en jeu. L'introduction de représentations conceptuelles est, en premier lieu, motivée par le désir, assez naturel, de faire figurer dans le modèle une contrepartie de la notion de signification lexicale qui, nous l'avons vu, est assumée par la propriété intentionnelle des concepts. Un deuxième intérêt d'introduire des représentations conceptuelles est de doter le modèle d'un substrat sur lequel les mécanismes de combinaison sémantique peuvent opérer. À l'agencement des mots dans une expression langagière peut correspondre une combinaison de significations lexicales qui permet de doter l'expression d'un sens. Ainsi, les représentations conceptuelles acquièrent un caractère compositionnel : elles entrent dans des combinaisons pour engendrer d'autres représentations plus complexes. La manière dont les représentations conceptuelles se combinent dépend de leur forme, et la forme du résultat est construite à partir de la forme des constituants. Ainsi, les concepts, de par leur forme, possèdent une propriété causale.

Une autre raison d'introduire des représentations conceptuelles pour expliquer le traitement cognitif du langage est liée au souci de reproduire les inférences qui sont faites spontanément par les sujets. Les représentations conceptuelles semblent présenter un rôle inférentiel : la représentation associée à une expression langagière peut déclencher d'autres représentations, par exemple par un processus de déduction. Là encore, il est raisonnable de tenter d'expliquer les inférences de manière causale à partir de la forme des représentations conceptuelles qui les déclenchent.

La forme des concepts peut donc être invoquée pour expliquer l'interface avec le langage et l'interface avec le raisonnement. Il est même possible de faire reposer sur la forme des concepts leur signification elle-même : pour certains auteurs, la caractérisation de la

forme des concepts est suffisamment contrainte par les interfaces avec le langage et le raisonnement, si bien que toute idée d'ancrage devient superflue. Dans ce cas, la signification des concepts, au moins en partie, est ramenée à leur seule propriété causale. Ce courant de pensée est qualifié d'internaliste. Dans sa version radicale, l'internalisme conceptuel ignore toute idée d'interface avec l'expérience, si bien que le système conceptuel apparaît comme un jeu dépourvu d'ancrage. Il existe cependant des versions modérées de l'internalisme conceptuel dans lesquelles la forme des représentations conceptuelles contribue partiellement à l'explication de l'intentionnalité. Si l'on peut encore parler de contenu dans le cas où l'on s'intéresse aux rapports que les concepts entretiennent entre eux, il s'agit d'un contenu étroit.

> The idea of the two–factor version is that there are two components to meaning, a conceptual role component that is entirely "in the head" (this is narrow meaning) and an external component that has to do with the relations between the representations in the head (with their conceptual roles) and the referents and/or truth conditions of these representations in the world. [...] The internal factor, conceptual role, is a matter of the causal role of the expression in reasoning and deliberation and, in general, in the way the expression combines and interacts with other expressions so as to mediate between sensory inputs and behavioral outputs. (BLOCK 1986 [7] p. 627)

Dans cette approche, qualifiée de sémantique du rôle conceptuel (*Conceptual Role Semantics*), le concept est déterminé par sa fonction au sein du système conceptuel, et c'est cette fonction qui lui procure, en partie, sa propriété intentionnelle. Il s'agit là d'un changement de perspective par rapport à la manière dont nous avons introduit la notion de concept. Nous avons caractérisé les concepts en tant que représentations dotées d'un contenu dans le domaine de l'expérience, avant de leur conférer une fonction au sein du système conceptuel. L'internalisme conceptuel, quant à lui, procède de manière inverse en caractérisant les concepts par le biais de la fonction qu'ils remplissent dans le comportement de l'individu qui les possède.

> A possession condition for a particular concept specifies a role that individuates that concept. The possession condition will mention the role of the concept in certain transitions that the thinker is willing to make. These will be transitions that involve complete propositional thoughts involving the concept. In some cases they are inferentional transitions; in others they are transitions from initial states involving perceptual experience. Normally, a possession condition has several clauses, each treating of a different kind of case. (PEACOCKE 1992 [81] p. 107)

Les concepts logiques servent de modèle à cette approche. Par exemple, le concept de conjonction peut être caractérisé par sa fonction dans la réalisation de certaines inférences par les individus qui le possèdent (PEACOCKE 1992 [81]). Cette fonction peut être représentée par l'ensemble de deux règles d'élimination ($P \wedge Q \vdash P$ et $P \wedge Q \vdash Q$) et une règle d'introduction ($P, Q \vdash P \wedge Q$). La possession du concept de conjonction est ainsi caractérisée d'une manière qui garantit la production de toutes les inférences ayant la forme de l'une des trois règles citées. Si l'on généralise cette méthode, il s'agit de déterminer, pour chaque concept, l'ensemble des clauses nécessaires qui cernent sa fonction cognitive. Cette fonction cognitive s'étend au pouvoir d'évocation du concept qui le lie aux mécanismes perceptifs ou effecteurs.

Certes, il nous parait naturel que le postulat des concepts s'accompagne d'une description exacte de la fonction qu'ils sont censés remplir dans le système cognitif humain. Cependant, nous pensons que, dans une approche représentationnaliste, cette tâche ne peut être entamée qu'en définissant, de manière claire, d'une part le domaine où les concepts

acquièrent leur contenu, et d'autre part les mécanismes qui agissent sur les concepts d'après leur forme. Dans notre schéma de base, le système conceptuel retient trois interfaces avec la perception, le langage et le raisonnement. Notre internalisme méthodologique exige que les concepts acquièrent leur contenu dans le domaine de l'expérience. Ainsi seule l'interface avec la perception fournit aux représentations conceptuelles l'ancrage nécessaire pour qu'ils possèdent une propriété sémantique. Les deux autres interfaces, quant à elles, permettent d'établir des liens associatifs, des liens de déclenchement par appariement ou des liens de contrôle, qui lient les mécanismes à l'œuvre dans le système conceptuel aux mécanismes du langage et du raisonnement. C'est l'explication de ces mécanismes qui nécessite l'analyse de la forme des représentations conceptuelles[5]. Or, selon l'hypothèse de l'internalisme conceptuel, l'ensemble des relations que les représentations conceptuelles peuvent entretenir avec d'autres entités mentales intervient dans la détermination de leur contenu. Notamment, les relations inter-conceptuelles prennent un rôle déterminant dans la définition du contenu étroit des concepts. Nous estimons qu'il s'agit ici d'un emploi métaphorique de la notion de contenu : dans un modèle représentationnaliste du système conceptuel, les relations inter-conceptuelles ne peuvent intervenir que dans la détermination de la structure interne des représentations conceptuelles ; ce n'est que d'une manière indirecte que cette structure peut être pertinente pour l'ancrage des représentations structurées[6]. Répétons que ce que nous appelons internalisme méthodologique n'est pas lié par l'exigence posée par l'internalisme conceptuel qui veut que le contenu des concepts soit déterminé, pour l'essentiel, de manière fonctionnelle. Notre position laisse par exemple la possibilité que chaque concept soit directement ancré dans l'expérience.

La notion de forme, en conférant un caractère fonctionnel aux représentations conceptuelles, réintroduit l'idée du sens des mots. La référence d'un mot est donnée par le contenu du concept correspondant. Son sens est donné, concrètement, par la forme de ce concept. La forme du concept précise la fonction du mot dans la communication langagière, c'est-à-dire les rapports qu'il entretient par son sens avec les autres expressions langagières. Le sens du mot pomme est ainsi représenté par la forme du concept associé qui lui donne la possibilité d'intervenir dans un syntagme comme une pomme rouge, ou dans une assertion comme la pomme est un fruit sucré.

À partir du moment où les concepts possèdent une forme, il faut se poser la question du rapport entre cette forme et celle de la représentation non conceptuelle qui l'évoque et qu'il évoque. On peut imaginer qu'elles soient de même nature, qu'elles soient qualitativement différentes mais que leurs structures se correspondent, ou enfin qu'il n'y ait aucun lien entre les deux. Dans les deux premiers cas, l'ancrage devient trivial. Dans le troisième cas, le problème de l'ancrage se révélera particulièrement délicat à résoudre.

3.2. Percepts intégrés

L'une des fonctions remarquables attribuées aux concepts est celle de la catégorisation. Les concepts nous permettent d'avoir un jugement sur les données de la perception, jugement qui est présenté comme binaire dans certains modèles. Ainsi, le concept LIVRE nous permet de catégoriser les objets du monde perçu de manière à en extraire, de manière plus ou moins précise, tout ce qui mérite d'être qualifié de livre.

[5] Nous développerons cette analyse en posant des questions sur, d'une part, le rôle inférentiel, et d'autre part, le caractère compositionnel des concepts (CF. CHAPITRES 4 & 5).

[6] Nous aurons l'occasion de reparler en détail des motifs et des méthodes pour structurer les concepts lexicaux, et des conséquences d'une telle entreprise (CF. CHAPITRES 7).

La manière la plus simple d'expliquer la capacité de catégorisation est sans conteste de l'imaginer comme une activité de comparaison. Ainsi, est une instance du concept LIVRE tout ce qui ressemble au livre typique. La caractéristique fondamentale de cette approche est que les deux éléments de la comparaison, le percept à catégoriser et le concept qui sert d'étalon sont de même nature. Son intérêt est de résoudre de la manière la plus simple qui soit la question de l'intentionnalité : non seulement le percept typique est le meilleur candidat pour constituer le contenu du concept, mais il n'y a plus de raison de distinguer le concept de son contenu. Les concepts, selon cette approche, ne sont rien d'autre que des percepts moyens, issus d'une intégration d'expériences multiples. Il s'agit donc, à la base, d'une théorie purement empirique des concepts.

Avantages de l'approche empiriste des concepts

L'intérêt premier de l'approche empiriste est de résoudre le problème de l'ancrage. Les concepts étant des percepts moyens, *id est* des invariants perceptifs, les concepts sont *ipso facto* ancrés dans l'expérience. Cette forme d'ancrage présente l'avantage d'être économique : contrairement aux théories qui postulent l'existence d'un espace conceptuel qualitativement distinct de celui des expériences, il n'y a pas, dans l'approche empiriste, de duplication des représentations. Ainsi, nous pouvons exprimer le fait qu'un zèbre possède des rayures parce que les rayures sont présentes dans la perception du zèbre. Alors qu'un système symbolique est obligé de dupliquer cette information sous une forme explicite, l'approche empiriste se contente d'exploiter l'information qui est disponible dans les données de la perception. Du point de vue de la modélisation, il s'agit là d'un avantage considérable, car toute duplication pose le problème de la complétude et de la cohérence.

L'ancrage empiriste résout un autre problème délicat, celui de l'acquisition. L'enfant acquiert de nouveaux concepts en généralisant ses percepts. Pour réaliser cette opération inductive, plusieurs mécanismes simples ont été proposés. Le mécanisme empiriste par excellence est un mécanisme statistique d'extraction de régularités. Ainsi, ne sont retenues dans le percept moyen que les caractéristiques qui sont fréquemment présentes dans les instances à partir desquelles il est construit. C'est ainsi que fonctionnent les réseaux de neurones artificiels classiques, qu'ils soient supervisés ou non.

> One of the basic principles which drives learning in neural networks is *similarity*. Similar inputs tend to yield similar outputs. Thus, if a network has leaned to classify a pattern, say 11110000, in a certain way then it will tend to classify a novel pattern, *e.g.*, 11110001, in a like fashion. Neural networks are thus a kind of analogy engine. The principle of similarity is what lets networks generalize their behaviors beyond the cases they have encountered during training. (ELMAN & BATES & JOHNSON & PARISI 1996 [32] p. 59)

Cette capacité de généralisation est à la base de la formation des percepts moyens. Le système extrait les régularités des exemples qui lui sont soumis. Les aspects contingents, qui varient d'un exemple à l'autre, ne laissent pas de trace, alors que les aspects invariants sont mémorisés ensemble. Une fois que cet apprentissage a eu lieu, le système peut reconnaître un objet qu'il n'a jamais perçu, à condition que cet objet possède un nombre suffisant de caractéristiques en commun avec l'un des ensembles invariants qui ont été mémorisés. Certains problèmes cognitifs ne possèdent pas cette propriété de continuité selon laquelle des entrées proches correspondent à des sorties proches. Un contre-exemple bien connu est celui de l'induction de la grammaticalité des énoncés du langage (CHOMSKY 1975 [15]). Néanmoins, la formation des catégories semble posséder, pour l'essentiel, cette propriété de continuité que les réseaux de neurones permettent de reproduire.

Noter que l'apprentissage statistique par extraction de régularités n'est pas le seul mécanisme permettant la formation de catégories. La théorie de la Gestalt et le constructivisme peuvent être présentés comme des approches fondées sur des mécanismes d'induction de "bonnes formes" à partir d'un ensemble réduit d'exemples (DESSALLES 1998 [29]). Dans ce cas, les concepts sont ces bonnes formes acquises par l'expérience qui servent ensuite à catégoriser les situations.

Le fait de ramener l'ancrage des concepts à un mécanisme de catégorisation possède plusieurs autres mérites. Le moindre n'est pas de pouvoir gérer la gradualité des jugements de catégorisation, ce que bien des systèmes symboliques sont incapables de faire. Ainsi, bien qu'un gland ne soit pas un chêne, le gland qui germe et qui pousse constitue chaque jour une meilleure instance du concept CHÊNE, sans qu'aucune discontinuité ne soit jamais franchie. De même, les liens de ressemblance qui peuvent unir les concepts, les airs de famille, sont naturellement pris en charge par la ressemblance des représentations dans l'espace perceptif. Ainsi, nous n'avons pas besoin d'explication pour percevoir la ressemblance entre une puce électronique et une vraie puce. Un tel appariement sera beaucoup plus délicat à effectuer pour des représentations conceptuelles ne comportant que des descriptions symboliques explicites.

Inconvénients de l'approche empiriste des concepts

Si, malgré le caractère impressionnant de ces points positifs, de très nombreux auteurs ont imaginé un système conceptuel disjoint du système perceptuel, c'est que les inconvénients de l'approche empiriste leur ont semblé rédhibitoires. Une partie de ces inconvénients est liée aux deux autres interfaces, laissées de côté par l'approche purement empiriste : l'interface avec le langage et l'interface avec le raisonnement.

Le principal problème que la formation empirique des concepts ne permet pas de résoudre est celui des phénomènes systématiques. Les concepts sont impliqués dans des opérations systématiques, que ce soit dans la catégorisation, les inférences ou la composition, qui sont incompatibles avec le caractère fondamentalement graduel inhérent aux mécanismes statistiques postulés dans l'approche empiriste. Par exemple, une entité inclue dans le concept VIVANT doit être systématiquement exclue du concept MORT. Or, pour un système empiriste, toute entité perçue est plus ou moins vivante, certaines franchement, d'autres presque pas, mais le lien entre l'entité perçue et le concept VIVANT n'est jamais strictement nul, car la statistique ménage toujours une ressemblance, même très faible, entre deux représentations quelconques. La frontière entre les concepts VIVANT et MORT n'est donc pas systématique comme il se devrait. De la même manière, la systématicité de l'inférence selon laquelle certaines entités perçues peuvent mourir dès lors qu'elles sont vivantes n'est pas assurée, car l'induction statistique dont elle dépend n'offre aucune garantie. Enfin, la systématicité de composition est également absente : rien ne garantit un sens systématique à une construction comme X est malade dès lors que la variable X désigne une entité vivante.

La nature statistique de l'apprentissage empiriste pose un deuxième problème. Un système empirique est peu sensible aux exceptions et ne leur associe pas d'explication. Dans un réseau de neurones classique, les exceptions n'étant pas représentatives de la classe, sont ignorées en tant que points aberrants de l'échantillon statistique. L'apprentissage d'une exception, par exemple une entité vivante immortelle, demande donc un renforcement répété, d'autant plus important que le lien conceptuel VIVANT - MORT a été fortement imprimé, jusqu'à ce que l'exception puisse constituer une classe à elle seule. Cet apprentissage ne produit jamais un résultat strict, et surtout mémorise l'exception sans pouvoir en stocker la raison. Or, un enfant humain qui lit une fiction comprend aisément, en une seule fois, que tel personnage peut traverser les âges sans vieillir parce qu'il possède un don particulier. Si son

fonctionnement conceptuel obéissait aux seules lois de l'apprentissage statistique empiriste, l'enfant serait incapable de comprendre ce type d'histoire.

Un autre grand problème rencontré par la conception empiriste du système conceptuel est la dépendance du système par rapport à son histoire. Pour que deux individus aient des concepts comparables, il faut qu'ils aient connu des expériences similaires en moyenne, qualitativement et quantitativement. Or, cette exigence est problématique dans les domaines où les individus n'ont eu accès qu'à des expériences limitées. Par exemple, un individu ne pourra pas communiquer à propos de ce qui est nouveau pour les autres. Il devra se limiter aux expériences communes, ce qui offre une piètre modélisation de la communication humaine. Pour expliquer que les enfants parviennent à des états de connaissances similaires malgré des expériences diverses et limitées, certains auteurs ont émis l'hypothèse que ces états de connaissances constituaient de "bonnes formes", dotés par exemple de la propriété de clôture opérationnelle, *id est* de fermeture par composition (PIAGET 1945 [83]). L'intérêt des bonnes formes est que leur apprentissage ne nécessite que peu d'exemples. Par exemple, le concept de justice, développé en plusieurs étapes par les enfants, présente de nombreuses symétries. L'enfant passe d'un état égo-centré, dans lequel est juste tout ce qui lui profite, à un état de justice réciproque dans lequel il considère comme injuste vis à vis des autres ce qu'il n'aimerait pas qu'on lui fasse (PIAGET 1932 [82]). Ce concept de justice, que l'on peut noter `juste(acteur , patient , acte)`, est invariant pour toutes les substitutions d'individus aux deux variables `acteur` et `patient`. Ceci inclut en particulier la possibilité d'adopter le point de vue de l'autre, ce dont le très jeune enfant est encore incapable. C'est parce que le concept adulte de justice présente de nombreuses symétries qu'il est atteint de manière identique par des enfants différents, soumis à des expériences différentes. Les concepts produits par généralisation statistique à partir de peu d'exemples constituent également des bonnes formes : la catégorie construite à partir de la généralisation des percepts est invariante pour toutes les transformations qui affectent les dimensions qui sont laissées de côté dans la généralisation. Ainsi, un concept empirique `POMME` peut ignorer la couleur précise de la pomme, sa taille, la présence de taches, *et cætera*. La catégorie créée à partir d'un percept moyen de pomme sera invariante pour toutes les transformations affectant ces aspects "anecdotiques".

Le fait que différents apprenants, exposé à des expériences limitées et différentes, convergent vers des bonnes formes n'est pas fortuit. Les mécanismes d'apprentissage de portée générale, comme les mécanismes empiristes ou constructivistes, permettant d'induire des catégories, ont la propriété d'être isotropes : ils sont insensibles à des changements de repère. Or, de tels mécanismes isotropes produisent des bonnes formes, *id est* des formes invariantes pour de nombreuses transformations (DESSALLES 1998 [29]). Dans le domaine conceptuel, il existe certainement des bonnes formes. Pour autant, on ne peut pas considérer que le système conceptuel se résume à un catalogue de bonnes formes. En particulier, les rapports qu'un concept donné entretient avec d'autres concepts ne présentent pas la belle symétrie qu'on est en droit d'attendre d'un système d'induction général. Par exemple, les relations conceptuelles évoquées au cours d'une conversation déterminent un ensemble de situations possibles de très faible symétrie (DESSALLES 1993 [28]). Si le seul mécanisme d'apprentissage disponible était un mécanisme inductif isotrope, empiriste ou constructiviste, la constitution des ensembles de connaissances que nous mobilisons dans les conversations serait inexpliquée. Cette situation reproduit, au niveau conceptuel, la difficulté d'expliquer l'acquisition des connaissances syntaxiques par des moyens empiriques ou constructivistes (CHOMSKY 1975 [15])[7].

[7] Nous aurons l'occasion de développer cette critique plus tard (CF. ANNEXE).

Le fait de considérer qu'il n'y a pas de différence de nature entre les percepts et les concepts, et que ceux-ci sont identifiables à des percepts moyens, est une idée à la fois très séduisante et très décevante. À la fin de ce travail, nous essaierons de montrer que cette idée et ses avantages peuvent être récupérés, à condition de compléter le dispositif inductif par un mécanisme dynamique de production de symboles. Auparavant, nous allons considérer en détail l'option opposée, selon laquelle les concepts sont d'une autre nature que les percepts.

3.3. Symboles

L'idée de symbole s'oppose par plusieurs aspects à la notion de percept intégré. Un percept intégré ne peut, au mieux, que représenter une catégorie d'objets perçus, alors qu'un symbole est avant tout un élément combinable ayant, par ailleurs, des propriétés représentationnelles.

> […] Nor can categorical representations yet be interpreted as « meaning » anything. It is true that they pick out the class of objects they « name », but the names do not have all the systematic properties of symbols and symbol systems […]. They are just an inert taxonomy. For systematicity it must be possible to combine and recombine them rulefully into propositions that can be semantically interpreted. « Horse » is so far just an arbitrary response that is reliably made in the presence of a certain category of objects. There is no justification for interpreting it holophrastically as meaning « This is a (member of the category) horse » when produced in the presence of a horse, because of the other expected systematic properties of « this » and « a » and the all–important « is » of predication are not exhibited by mere passive taxonomizing. (HARNAD 1990 [46] p. 343)

La tradition empiriste postule des représentations conceptuelles qui entretiennent des rapports de ressemblance avec les données de la perception. La topologie, voire la métrique de ressemblance, au sein de l'espace des concepts, est la même que dans l'espace des percepts. Dans beaucoup de modèles, ceci est dû au fait qu'il s'agit du même espace. Cette situation contraste avec les autres modèles, dans lesquels ce lien de ressemblance n'existe pas. Lorsque l'on met en avant le caractère symbolique des concepts, que l'on dote les concepts de propriétés formelles pour permettre leur inclusion dans des calculs dictés par les mécanismes grammaticaux ou déductifs, les représentations obtenues perdent nécessairement tout lien de ressemblance et toute relation topologique ou métrique avec leur contenu. Cette perte, qui se constate dans les différents modèles qui ont pu être proposés, est due au fait que les mécanismes généraux qui conservent les relations de distance et de voisinage possèdent la propriété d'isotropie, alors que les mécanismes compositionnels qui engendrent les représentations conceptuelles complexes en sont dépourvus (DESSALLES 1998 [29]).

Si aucune ressemblance n'existe entre les concepts et les percepts, si l'organisation des concepts ne reproduit en rien l'organisation des percepts, alors l'ancrage des concepts doit être le fruit d'un "apprentissage", soit ontogénétique, soit phylogénétique. Or l'établissement de ce lien d'ancrage pose le problème de la connexion. Par quel mécanisme les concepts, qui sont ici des entités symboliques, peuvent-ils recevoir un contenu dans l'espace des percepts ? On comprend aisément comment un mot *M* peut se retrouver associé à un percept *P*. Il suffit que *M* soit prononcé en présence du sujet lorsque celui-ci perçoit *P*. Si le mot *M* doit renvoyer en même temps au concept *C*, il faut déterminer le mécanisme par lequel le triangle *M* □*P* □*C* se retrouve complété. Or *C*, en tant qu'entité symbolique propre au système conceptuel, est en quelque sorte hors d'atteinte.

Ce problème de la connexion concerne toutes les théories symboliques des concepts. Certes, certaines théories parviennent à dériver l'ancrage des concepts de l'ancrage de leurs composants. Ainsi, si le concept TULIPE est défini, au sein du système conceptuel, à l'aide du concept FLEUR, il hérite une partie de son ancrage de ce dernier. Cette méthode peut, en principe, résoudre le problème de l'ancrage pour la plupart des concepts. Il n'en reste pas moins que la connexion restera inexpliquée pour certains concepts, dits primitifs, à partir desquels tous les autres sont construits.

Nous allons considérer successivement deux types d'explication de l'établissement de l'intentionnalité, invoquant respectivement la communication et l'information. Au départ, ces explications ont été produites dans un contexte externaliste, pour expliquer que les concepts se trouvent reliés à des entités du monde. Dans chaque cas, nous examinerons comment ces cadres théoriques peuvent être transposés dans notre propre cadre, en prenant l'explication de l'intentionnalité comme un mécanisme d'ancrage. Même dans l'hypothèse d'un monde objectif, notre expérience reste le seul lien que nous pouvons avoir avec ce monde. L'intentionnalité des concepts, comprise dans ces théories comme le lien qui unit ces concepts au monde, vient donc nécessairement de la composition de deux mécanismes, l'intégration de l'expérience perceptive et l'ancrage.

3.4. Intentionnalité et communication

Certaines approches n'abordent pas la question de l'intentionnalité comme un problème de connexion entre des représentations mentales, mais comme un problème d'association entre des mots et des catégories du monde. Pour de telles approches, qui adoptent une perspective externaliste, il est essentiel d'expliquer comment le langage parvient à décrire le monde de manière "correcte". Comment un individu donné parvient-il à connaître l'ensemble des liens mot - référence, même pour des entités du monde qu'il n'a jamais rencontrées ? La réponse suppose que le lien causal entre le mot et sa référence a généralement une origine qui dépasse l'histoire de l'individu. Si je désigne un objet précis de mon champ visuel en prononçant le mot arbre et que mon interlocuteur saisit mon intention, c'est parce que mon interlocuteur a appris ce mot quand il était enfant et que je l'ai moi-même appris en cours de français quand mon professeur désignait une image dans le livre de cours. Pour que je puisse communiquer le contenu du mot arbre, ce contenu a dû m'être communiqué. Il a été aussi communiqué à mon professeur, et ainsi de suite le long d'une chaîne d'usage jusqu'à la première fois où un individu a glosé l'objet par le mot. La communication assure le lien entre le mot et sa référence et le propage.

Le cas typique, qui sert de modèle à cette approche, est celui des noms propres. Les noms propres sont le résultat d'un acte de baptême. En donnant un nom à un individu, la communauté décide de le désigner par une étiquette qui glosera désormais l'enfant pour ceux qui le connaissent. Le nom constitue alors le moyen le plus simple pour pouvoir communiquer verbalement à propos de cet individu. Grâce à cette étiquette conventionnelle, on peut même parler de l'individu avec quelqu'un qui ne le connaît pas. La personne sait que l'étiquette ne fait que remplacer sa référence.

> Someone, let's say a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain. A speaker who is on the far end of this chain, who has heard about, say Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can't remember from whom he first heard about Feynman or from whom he ever heard of Feynman. (KRIPKE 1972 [62] p. 91)

L'élargissement de la métaphore de baptême initial aux noms communs semble assez intuitif quand il s'agit des espèces naturelles ou des artefacts avec lesquelles les individus sont en interaction directe. On peut imaginer que l'on apprend des mots comme chat ou voiture quand ils sont utilisés pour désigner des objets de la vie quotidienne. Grâce à la chaîne communicationnelle, il n'est même pas obligatoire que l'individu ait eu une expérience directe des objets désignés, si bien que la communication par ostension n'est pas nécessaire pour l'apprentissage individuel.

> [...] the species–name may be passed from link to link, exactly as in the case of proper names, so that many who have seen little or no gold can still use the term. Their reference is determined by a causal (historical) chain [...]. (KRIPKE 1972 [62] p. 139)

Si les individus ne peuvent pas profiter de l'ostension pour établir le lien mot - référence, il existe un autre moyen, la description, par lequel ils accèdent au moyen d'utiliser le mot dans les circonstances adéquates. Dans le cas des entités abstraites ou imaginaires désignées par des mots comme bourse ou Pégase, le baptême initial est plus difficile à concevoir. Il suffit dans ce cas d'imaginer que la description sert non seulement à contraindre le lien mot - référence pour tous les individus de la chaîne de communication, mais qu'elle sert également à effectuer le "baptême" de l'entité pour la communauté (KRIPKE 1972 [62]). Ainsi, la première description qui a été faite de Pégase en tant que cheval ailé a permis d'établir un lien mot - référence qui s'est ensuite propagée d'individu en individu, si bien que le mot Pégase, sans être équivalent à la description qui lui est associée, possède une référence pour tous les individus de la communauté.

Cette approche de la notion de contenu repose sur deux considérations de base. La première est liée au caractère indexical d'une grande partie des expressions langagières (PUTNAM 1975 [87]). Selon cette idée, la fonction principale du mot arbre est de désigner sa référence. La seconde est d'ordre social : le fait que l'on utilise le terme arbre et non un autre terme pour désigner sa référence est le résultat d'un accord dans la communauté de communication (PUTNAM 1975 [87]).

La question qui nous importe, dans ce travail, est de savoir comment une telle théorie du contenu peut s'interpréter dans une approche cognitive représentationnaliste du langage naturel. Notre but n'est pas de savoir comment les individus parviennent à employer les mots d'une manière objectivement correcte, mais plutôt de rendre compte de la capacité des individus à comprendre les énoncés langagiers de manière à effectuer certaines inférences et à produire d'autres énoncés appropriés. Dans un cadre représentationnaliste, le mot, qui n'est autre qu'une forme phonologique propre à la langue parlée dans la communauté, peut évoquer une représentation mentale chez les locuteurs de cette langue. C'est cette représentation mentale, le concept, qui déclenche les processus compositionnels et inférentiels. Pour que cela soit possible, il faut que le concept soit lié, à la suite d'un apprentissage, au mot qui l'évoque. Si l'on accepte que le concept est lié à un certain type de perception, par exemple la perception d'un lien causal entre un phénomène et ses effets, alors il suffit que le mot soit associé une fois à cette situation causale pour qu'il se retrouve associé au concept pour tous les individus de la communauté linguistique.

> They could all use the term "electricity" […] what they do have in common is this: that each of them is connected by a certain kind of causal chain to a situation in which a *description* of electricity is given, and generally a *causal* description – that is, one which singles out electricity as the physical magnitude *responsible* for certain effects in a certain way. […] Let us call this event – my acquiring the ability to use the term "electricity" in this way – an *introducing event*. It is clear that each of my later uses will be causally connected to this introducing event, as long as those uses exemplify the ability I required in that introducing event. (PUTNAM 1975 [87] p. 199)

De cette manière, le lien mot - concept peut s'expliquer, au même titre que le lien mot - objet considéré par ces auteurs, par une chaîne historique causale qui associe, à chacun de ses maillons, une représentation phonologique et une représentation conceptuelle. Le point essentiel que l'on peut retenir de cette approche historique et conventionnelle de la signification est que le fait que les mots aient un sens suppose que les individus interagissent de manière causale avec leur environnement et interagissent de manière causale entre eux, par le biais de la communication.

Cette manière de concevoir la création de nouvelles significations et leur propagation laisse une part minime à l'apprentissage et aux difficultés qu'il présente. Le problème est de rendre plausible le fait que deux individus finissent par accorder la même signification au même mot. Même dans le cas des noms propres, la réponse ne va pas de soi. Lors de l'événement baptismal, les personnes présentes ont l'occasion d'associer le nom non pas à l'enfant lui-même, en tant qu'entité objective, mais à une perception qu'ils en ont. Il faut qu'un mécanisme leur permette de généraliser le nom à l'individu quel que soit son âge, leur évite d'associer le nom à l'enfant seulement lorsqu'il est dans son landau, *et cætera*. La situation est encore plus problématique dans le cas des espèces naturelles. À partir du moment où l'on renonce à la simplification qui consiste à penser que les entités du monde nous apparaissent directement, détachées de toutes les autres, l'établissement des associations mots - concepts, médiatisées par les associations mots - percepts, devient moins évident. Il s'agit d'expliquer ce qui nous permet de réemployer le même mot chat dans des situations différentes dans lesquelles le contexte d'apparition de l'animal peut varier du tout au tout. Il faut postuler l'existence de capacités inductives puissantes qui permettent à deux individus, ou à un même individu dans deux situations différentes, de repérer des éléments communs qui provoquent l'utilisation appropriée d'un même mot. L'explication du lien historique causal de dénomination y perd la simplicité qui faisait son attrait. S'il existe un mécanisme psychologique complexe par lequel les individus parviennent à déterminer la référence d'un mot en contexte, le fait que ce mot soit prononcé en leur présence dans des circonstances appropriées, bien que nécessaire, ne joue qu'un rôle marginal dans l'établissement de la signification.

Il est également difficile de se contenter d'invoquer l'histoire des interactions lorsqu'il s'agit d'expliquer l'établissement du sens des mots, notamment les mots abstraits, dont la signification est obtenue par des descriptions utilisant d'autres mots. L'interprétation des descriptions suppose là encore des mécanismes riches, laissant à l'origine historique des associations mot - objet une part anecdotique. L'importance des mécanismes cognitifs est encore plus flagrante lorsque la description sert à désigner une situation dont l'individu est censé inférer le concept décrit. Ainsi, une description comme ce qui surnage lorsqu'on fait cuire du lait suppose des capacités à constituer un concept à partir d'une situation imaginée ou remémorée. Tout sujet qui apprend un mot nouveau pour lui est confronté à un problème de même difficulté que celui qu'ont connu ses ancêtres lors de la création du lien mot - objet dans la communauté.

La question de l'émergence de significations dans une communauté a été explorée expérimentalement à l'aide de simulations (KAPLAN 2001 [57]). Il apparaît que l'acte initial de dénomination d'un objet perçu dans l'environnement par un agent ne suffit pas à l'acquisition du concept par les autres agents. Il faut que ceux-ci aient une perception comparable, puis affinent leurs catégories lors des utilisations ultérieures du même mot. Les agents parviennent ainsi à former des catégories conceptuelles pertinentes en affinant des divisions selon une procédure dichotomique, par exemple sur la couleur ou la forme des objets, de manière à discriminer les objets présents. Il peut en résulter des dialogues de sourds : tel objet sera signalé par le mot *gorewa* par l'agent 1, ce qui, pour lui, signifie "rouge" ; l'agent 2, entendant *gorewa*, comprendra par exemple "triangle". La communication peut réussir si le seul objet rouge se trouve être triangulaire. Heureusement, confrontés à d'autres situations, l'un des deux agents au moins sera amené à réviser son lexique.

Ce type d'expérience montre que l'histoire de la communication entre agents, qui ressemble plus à un réseau complexe d'interactions qu'à une simple chaîne de communication, se révèle déterminante pour l'adoption d'un mot plutôt qu'un autre. L'histoire des interactions en contexte permet donc, en principe, d'expliquer l'émergence d'un ensemble donné de concepts dans une communauté. En revanche, il serait illusoire de vouloir fonder une théorie des concepts en se limitant à l'historique des interactions. Les concepts que forment les robots préexistent en puissance dans leur capacité à affiner leurs catégories et à les associer avec des mots. L'observateur pourrait être tenté de penser que la signification des mots qu'ils emploient se situe dans le monde simplifié qui s'offre à leur caméra. Ainsi, le mot *gorewa* semble désigner le triangle rouge sur le tableau. Pourtant, ces robots diffèrent suffisamment de nous pour que nous soyons obligés de reconnaître qu'il n'en est rien. La signification de leurs mots se situe dans leur mémoire, dans les structures qu'ils ont affinées pour répondre aux besoins de la communication qui joue davantage le rôle de déclencheur des mécanismes internes qu'un rôle structurant. Ces robots, pas plus que les humains, ne sont télépathes. Ils ne peuvent pas se transmettre leurs catégories conceptuelles. Ils ne peuvent qu'échanger des mots en situation et employer leurs mécanismes internes pour tenter de parvenir à des catégories compatibles avec leur perception.

Il est indéniable que l'association entre les mots et les perceptions ne doit rien, la plupart du temps, à la forme phonologique des premiers. Les associations mots - percepts ne sont donc possibles que si certains individus les créent et les propagent. Cependant, comme les situations ne se présentent jamais deux fois sous le même aspect, le lien à expliquer est celui qui s'instaure entre le mot et le concept, ce dernier permettant de discriminer les percepts. L'étude des interactions entre individus ne permet pas de faire l'économie d'un mécanisme inductif par lequel ces individus parviennent à forger des concepts compatibles avec l'usage des mots par leurs partenaires.

3.5. Intentionnalité et information

L'histoire des interactions langagières ne peut suffire à rendre compte de l'intentionnalité des concepts : elle ne peut qu'expliquer pourquoi ce sont certains percepts qui sont nommés dans la communauté plutôt que d'autres. Si l'on s'intéresse au résultat de ce processus historique pour l'individu, force est de constater que celui-ci a acquis une capacité lui permettant d'employer les mots à bon escient. Une manière minimale de décrire cette capacité est de la présenter comme une disposition comportementale.

> [...] a stimulation σ belongs to the affirmative stimulus meaning of a sentence S for a given speaker if and only if there is a stimulation σ′ such that if the speaker were given σ′, then were asked S, then were given σ, then were asked S again, he would dissent the first time and assent the second. [...] What now of the strong conditional, the "would" in our definition of stimulus meaning? [...] what the conditional defines is a disposition, in this case a disposition to assent to or to dissent from S when variously stimulated.
>
> (QUINE 1960 [90] p. 32)

La présence de ces dispositions chez les individus est le résultat d'un apprentissage. Dans la doctrine behavioriste, le mécanisme d'apprentissage par lequel on acquiert de nouvelles significations ne peut être qu'un renforcement positif ou négatif.

> It remains clear in any event that the child's early learning of a verbal response depends on society's reinforcement of the response in association with the stimulations that merit the response, from the society's point of view, and society's discouragement of it otherwise.
>
> (QUINE 1960 [90] p. 82)

Pour ce genre d'approche, le fait de posséder des concepts n'est rien d'autre qu'une disposition comportementale conditionnée. On ne peut donc pas parler de contenu pour les concepts. Toute notion même de sémantique est exclue d'un tel schéma (QUINE 1960 [90]). Si l'on est représentationnaliste et que l'on fait découler les dispositions comportementales langagières de la possession de représentations conceptuelles, le seul point à retenir d'une telle théorie est le lien statistique, fruit d'un renforcement répété, qu'elle postule entre les concepts et les percepts.

La reconnaissance de ce lien statistique empêche de ne voir dans les concepts que le simple résultat d'un acte de désignation propagé à travers l'histoire des usages. Selon cette hypothèse, les sujets sont capables d'identifier l'objet qu'on leur désigne et de le reconnaître dans les situations ultérieures parce qu'ils ont forgé un concept qui leur permet de le discriminer (DRETSKE 1981 [31]). Cette fonction de discrimination justifie, d'un point de vue théorique, l'existence de concepts dotés d'un contenu. Le mot désigne un objet, mais cet objet n'existe perceptivement que parce qu'un concept le discrimine. Dans ce genre d'approche où, rappelons-le, les concepts ne sont pas de même nature que les percepts, le pouvoir de discrimination des concepts est attribué au fait que le concept co-varie avec l'objet perçu (DRETSKE 1981 [31]). Cette co-variation a la vertu d'expliquer également l'acquisition de ce pouvoir de discrimination.

> In teaching someone the concept *red*, we show the pupil variously colored objects at reasonably close range and under normal illumination. That is, we exhibit the colored objects under conditions in which *information* about their color is transmitted, received and (hopefully) perceptually encoded. [...] If the subject is to acquire the concept *red*, he or she must not only be shown red things (and, presumably, non–red things), they must be allowed to receive the information that they are red (and not red). The reason for this should be clear: it is the information that the object is red that is needed to shape the internal structure that will eventually qualify as the subject's concept *red*. We need information to manufacture meaning (the concept) because information is required to crystallize a type of structure with the appropriate semantic content. (DRETSKE 1981 [31] p. 194)

Non seulement la co-variation entre le concept et l'objet permet au sujet de former le concept puis, grâce à ce concept, de discriminer les occurrences futures de l'objet, mais elle explique également le fait que le contenu d'un concept reste stable au cours du temps dans la communauté. Ainsi, dans ce type de théories, le lien causal qui unit le concept à son contenu

est de nature statistique, attribué à une co-variation repérée par les sujets entre certaines de leurs perceptions et certains de leurs états mentaux.

Le principe de co-variation remet en cause l'idée de contenu purement conventionnel. Une signification qui serait un pur produit d'une convention sociale, propagée d'individu en individu, pourrait être totalement arbitraire. Le lien qui unit le concept à son contenu, dans la mesure où il repose sur une co-variation statistique, n'est pas arbitraire. Il prend un caractère naturel. Le concept, parce qu'il co-varie avec les objets perçus, devient porteur d'information. Pour un externaliste, les conséquences semblent impressionnantes. Les concepts, par leur intentionnalité, apportent une information sur le monde. On rejoint ainsi l'idée selon laquelle le langage nous renseigne sur la réalité qui nous entoure. Malheureusement, la connaissance que l'on peut espérer avoir du monde en étudiant les concepts est limitée en raison du manque de robustesse d'un lien intentionnel de nature statistique. Dans la mesure où l'erreur est possible, où la perception d'un chat dans l'obscurité peut conduire à tort à l'évocation du concept CHIEN, l'activation d'un concept n'offre aucune certitude sur l'état du monde. Si de telles erreurs sont fréquentes, on peut arriver à ce que le concept soit évoqué dans des conditions considérées comme impropres du point de vue externaliste : le concept CHIEN co-variera avec un ensemble hétérogène d'entités et non avec la catégorie "chien", considérée comme objective. Il semble n'y avoir aucune garantie théorique sur ce que l'évocation d'un concept peut signifier sur l'état du monde. Cette situation, qui est sans conséquences si l'on se place dans le cadre de l'internalisme méthodologique, semble inacceptable pour les tenants de l'externalisme.

La question de la validité du lien intentionnel, considéré comme informationnel, a conduit bon nombre d'auteurs à rechercher une distinction formelle entre l'évocation "normales" du concept et les conditions d'évocations considérées comme erronées. Ainsi, ce n'est pas parce que, dans des cas de perception anormale, le concept CHIEN viendra à être évoqué que ce concept recouvrera la signification "chien ou grand chat dans le noir". Même si le concept CHIEN co-varie avec des situations du type "grand chat dans le noir", cette co-variation peut être éliminée du lien intentionnel dès lors qu'on est capable de filtrer les situations anormales d'évocation. La question évidente est alors de savoir selon quel critère on peut distinguer les conditions normales des conditions anormales d'évocation. On peut considérer qu'il existe une asymétrie fondamentale qui alloue une prééminence aux conditions normales.

La nature de cette asymétrie est difficile à expliciter, en dehors du constat *a posteriori* que, comme les concepts sont porteurs d'information, il faut bien que les conditions normales d'évocation l'emportent sur les conditions anormales. On peut imaginer, par exemple, en raison de la nature statistique du lien d'évocation, que les conditions anormales sont plus rares. Cette supposition peut reposer sur une présomption d'efficacité du système conceptuel qui est capable de filtrer, dans la plupart des cas, les conditions adéquates d'évocation des concepts. Les concepts sont supposés avoir une fonction cognitive propre. On peut donc s'appuyer sur cette fonction cognitive pour affirmer que, dans des conditions normales de fonctionnement, les concepts ont un lien intentionnel non ambigu. Certains auteurs justifient cette idée de fonctionnement normal non seulement sur le plan cognitif, mais également dans une perspective phylogénétique (MILLIKAN 1984 [73]). Ils en appellent à la sélection naturelle pour fonder l'idée de fonction propre pour un concept.

Si nos concepts existent, c'est parce qu'ils sont efficaces, et ils sont efficaces parce que nous descendons d'individus qui ont été sélectionnés pour les avoir possédés, ou tout du moins pour avoir possédé les moyens de les acquérir en tant que concepts efficaces. Le contenu de tels concepts, forgés par la sélection naturelle, est celui pour lequel ils ont été sélectionnés dans la phylogenèse ou dans l'ontogenèse, non les configurations qui peuvent par moments les évoquer de manière erronée. Il semble que l'on récupère, par le biais de cette

fonction propre, une notion de contenu objectif pour les concepts. Un exemple souvent cité est celui de la perception d'une mouche par la grenouille. L'état neuronal provoqué par la vue de la mouche déclenche une réaction de prédation. Dans certaines conditions, le même comportement de prédation peut être provoqué par un leurre ou par un petit objet inerte emporté par le vent. Si l'on suppose que la réaction de prédation est adaptée pour la fonction biologique de prise d'aliment, on est en droit de dire que la mouche est la cause normale de la réaction de prédation et que le leurre en est une cause anormale. Dans ces conditions, la réaction de prédation peut être considérée comme l'indice de la présence d'une mouche, même dans le cas d'une erreur.

L'intérêt de ce type de raisonnement téléologique pour la compréhension du système conceptuel humain doit être relativisé. Notons tout d'abord que l'appel à la sélection naturelle pour justifier l'efficacité des concepts est critiquable. On peut l'accepter lorsqu'il s'agit d'une caractéristique biologique, quoique dans certaines limites (GOULD & LEWONTIN 1979 [45]). Par exemple, on peut imaginer qu'il existe un petit nombre de concepts primitifs innés qui sont de bons candidats pour être le fruit d'une évolution biologique. Or, cette éventualité est loin d'aller de soi, en particulier parce que nous parlons ici de l'innéité de concepts et non de l'innéité de sensations, comme nous aurons l'occasion de le souligner plus loin (CF. CHAPITRE 8). Lorsqu'il s'agit d'expliquer la genèse de l'ensemble des concepts, considérés comme appris au cours de l'ontogenèse, nous n'avons plus affaire à des caractéristiques biologiques à proprement parler, et l'argument de la sélection naturelle ne porte pas. Il faut donc se rabattre sur l'idée métaphorique d'une sélection ontogénétique analogue à la sélection darwinienne qui se ramène aux mécanismes d'apprentissage qui ont été proposés.

Par ailleurs, si l'appel à la sélection naturelle, dans le cas de l'exemple de la grenouille, apparaît comme justifié, dans le cas des concepts est moins évident. La fonction propre du comportement de prédation observé chez la grenouille ne fait pas mystère. Comment définir une fonction propre pour le concept CHIEN ? Si l'on dit qu'il s'agit de discriminer les chiens et que le concept CHIEN est biologiquement efficace pour cette fonction et aucune autre, le raisonnement est circulaire. On définit la fonction propre du concept CHIEN à partir de la catégorie "chiens", considérée comme une donnée objective, alors que l'objectif était, à l'inverse, d'apprendre quelque chose sur le monde à partir du concept. La seule chose qui peut être dite est donc que le concept doit avoir un contenu dans le monde, qui correspond à ce qu'il discrimine dans les conditions normales, sans qu'il soit possible d'en dire davantage puisque nous sommes dans l'impossibilité de définir ce qu'est le fonctionnement normal d'un concept donné. C'est d'ailleurs cette impossibilité fondamentale de faire une différence *a priori* entre discrimination réussie et discrimination erronée qui rend le projet externaliste si difficile à transposer à la modélisation cognitive.

Une tentative pour restaurer la cohérence de la vision externaliste est de renoncer à la fonder sur le seul critère informationnel. Si l'on reconnaît l'impossibilité, dans la plupart des cas, de caractériser de manière objective une différence entre les causes idéales d'évocation du concept et les causes erronées, il est préférable de poser que le contenu d'un concept est constitué par ce qui est commun à toutes les situations qui le déclenchent (FODOR 1990 [35]). De cette manière, le concept véhicule de l'information sur toutes ces causes. La conséquence, pour un externaliste, est que l'information portée par le concept ne suffit pas, à elle seule, pour définir son contenu. Si le concept CHIEN est déclenché dans certaines conditions par un chat, alors le chat dans ces conditions fait partie de l'ensemble des situations avec lesquelles le concept CHIEN co-varie, sans faire partie de son contenu. Que peut-on dire, alors, sur le contenu externe du concept ? Écrivons que le concept C peut être causé par une propriété P du monde dans les conditions s : $C = f(a_s(P))$. La fonction $a_s(P)$ désigne par exemple l'apparence de P dans les conditions s. Le même concept peut être causé par une propriété P' dans les

conditions s' : $C = f(a_{s'}(P'))$. Si la symétrie entre ces deux formules est parfaite, on ne peut espérer distinguer les conditions normales, représentées ici par s, des conditions anormales désignées ici par s'. Cependant, si l'on peut écrire que $a_{s'}(P) = a_{s'}(P')$ sans que $a_s(P) = a_s(P')$, alors nous tenons une dissymétrie qui permet de réhabiliter la notion de contenu P pour le concept C. Ainsi, nous pouvons affirmer $C = f(a_{s'}(P)) = f(a_{s'}(P'))$, mais $C = f(a_s(P)) \neq f(a_s(P'))$: P et P' deviennent discernables, en théorie, par rapport à C. L'écriture $a_{s'}(P) = a_{s'}(P')$ peut signifier que P et P' ont la même apparence lorsque la perception est sujette à un certain niveau d'imprécision s', ou qu'elles évoquent la même image si on tolère un niveau d'association s'. Ainsi, un chat peut passer pour un chien dans le noir, et le lait peut faire penser à une vache si on se laisse aller à faire des associations. L'asymétrie entre P et P' vis-à-vis de C permet de penser que l'une constitue le contenu de C et l'autre non. Si C porte de l'information sur P', c'est seulement parce que C porte de l'information sur P.

> Cows cause "cow" tokens, and (let's suppose) cats cause "cow" tokens. But "cow" means *cow* and not *cat* or *cow or cat* because *there being cat–caused* "*cow*" *tokens depends on there being cow–caused* "*cow*" *tokens, but not the other way around*. "Cow" means *cow* because, […] non–cow–caused "cow" tokens are *asymmetrically dependent upon* cow–caused "cow" tokens. "Cow" means *cow* because *but that* "*cow*" *tokens carry information about cows, they wouldn't carry information about anything*. (FODOR 1990 [35] p. 91)

Un externaliste, selon cette théorie, peut prolonger dans le monde le caractère asymétrique des évocations pour en conclure que certaines situations, parmi celles qui évoquent le concept, en constituent le contenu, alors que d'autres, qui l'évoquent tout autant, en sont exclues. Cette asymétrie ne doit rien au critère informationnel, qu'il soit phylogénétique ou ontogénétique, puisque toutes ces situations co-varient de la même manière avec le concept.

Dans le cadre de l'internalisme méthodologique au sein duquel nous nous situons, ce débat concernant le caractère robuste du contenu conceptuel perd beaucoup de son enjeu. La question devient de savoir si les concepts, considérés ici comme des états mentaux distincts de la perception, renvoient de manière non ambiguë à certains percepts, autrement dit aux percepts qu'ils sont censés normalement discriminer. L'argument téléologique revient alors à fonder cette normalité sur le fait que chez des êtres qui sont le produit de la sélection naturelle, le fonctionnement normal est celui qui a été sélectionné. Comme dans le cas externaliste, cela ne nous renseigne pas beaucoup sur le contenu du concept, car en général nous n'avons pas de moyen indépendant d'estimer l'incidence de telle ou telle discrimination sur la survie des individus qui en sont capables. L'argument de la dépendance asymétrique, quant à lui, se limite à postuler une distinction direct/indirect dans les conditions d'évocation des concepts, sans qu'il soit possible d'en dire plus sans une information indépendante sur les catégories que le concept est censé discriminer.

Notons que pour l'internalisme méthodologique, le lien entre les percepts et les concepts est nécessairement sujet à l'erreur. On ne peut pas appliquer l'espace continu des percepts sur l'espace symbolique des concepts sans qu'il se produise des distorsions. On rejoint ainsi les notions de détection en présence de bruit développées dans la théorie mathématique de l'information. La qualité de la discrimination d'un percept par un concept peut s'estimer par la mesure d'une certaine probabilité d'erreur. On peut même estimer la qualité d'un ensemble de concepts comme sa capacité à produire des discriminations qui minimise la distorsion des distances entre percepts.

Avec cette manière de concevoir l'opération de discrimination conceptuelle, l'internalisme méthodologique permet d'abandonner sans dommage l'objectif qui consiste à vouloir établir une "bijection" entre les concepts et un ensemble prédéfini de contenus

possibles. Comme suggéré plus haut, on peut définir le contenu d'un concept comme ce qui est commun à toutes les situations qui le déclenchent. De cette façon, on conserve l'idée de lien informationnel entre le concept et son contenu, puisque la co-variation est systématique. Cette définition a également le mérite de supprimer l'essentiel de ce qui constituait des "erreurs" dans le cas bijectif. Cet avantage est dû au fait que, comme dans le cas empiriste où les concepts étaient constitués de percepts généralisés, le lien intentionnel permet de respecter la gradualité inhérente à la perception. Lorsque l'on cherche une bijection entre les concepts et les catégories du monde, il faut être capable de donner avec précision la limite de la catégorie "chiens" hors de laquelle le concept CHIEN cesse d'être activé. Une telle gageure n'est pas exigée si le concept, bien qu'étant considéré comme une entité symbolique distincte des percepts, renvoie à un contenu correspondant à un percept moyen.

Un autre avantage de l'abandon du lien bijectif au profit de concepts à contenu moyen est de permettre d'envisager plus simplement un mécanisme d'apprentissage. Dans le cas externaliste, l'origine de l'adéquation parfaite entre les entités mentales que sont les concepts et les catégories du monde reste inexpliquée. Comme nous l'avons vu, l'appel à la sélection naturelle ne remplit pas le vide théorique dû à l'absence de critère indépendant pour décider ce qui, dans le monde, mérite, d'un point de vue biologique, d'être isolé comme le contenu d'un concept. La conséquence est un raisonnement circulaire dans lequel ce sont précisément les concepts que nous formons qui sont supposer favoriser la survie. Si, dans le cadre de l'internalisme méthodologique, nous considérons que le contenu des concepts se trouve dans le domaine des expériences, alors l'établissement de ce contenu au cours de l'ontogenèse semble aisé à expliquer. Il suffit d'invoquer un mécanisme statistique d'extraction de régularités. Cette facilité n'est malheureusement qu'apparente, comme nous allons le constater maintenant.

3.6. Le paradoxe de la connexion

Les différentes théories du système conceptuel humain ne sont pas égales devant le problème de la connexion du concept à son contenu. Pour les empiristes, le problème n'existe pas, car les concepts sont de la même nature que les percepts et le lien intentionnel est une identité. Les mécanismes de généralisation ou de construction de bonnes formes produisent des concepts qui, restant dans l'espace des percepts, ne cessent jamais d'être identiques à leur contenu.

Le problème de la connexion se pose de manière extrême, au point de conduire à un paradoxe, dès que les concepts sont considérés comme étant d'une autre nature que les percepts.

> […] there is a further constraint that whatever theory of concepts we settle on should satisfy: it must explain why there is so generally a content relation between the experience that eventuates in concept attainment and the concept that the experience eventuates in attaining. (FODOR 1998 [37] p. 132)

La différence de nature entre les représentations perceptuelles et les représentations conceptuelles rend le problème particulièrement délicat. Les concepts, nous l'avons dit, sont supposés être des représentations discrètes et digitales. Ils n'offrent pas la gradualité qui existe dans l'espace des percepts. Ils ne représentent pas non plus les relations de voisinage qui existent dans le domaine des percepts. Le mécanisme qui les connecte aux percepts ne saurait résulter d'un simple processus d'échantillonnage. Le lien qui semble exister entre

certains concepts et des prototypes perceptuels, loin d'être une évidence, est précisément ce qui requiert une explication.

> […] what doorknobs have in common qua doorknobs *is being the kind of thing that our kind of minds (do or would) lock to from experience with instances of the doorknob stereotype*. […] We have the kinds of minds that often acquire *the concept X from experiences whose intentional objects are properties belonging to the X–stereotype*. […] *Stereotype* is a *statistical* notion. The only theoretically interesting connection between *being a doorknob* and *satisfying the doorknob stereotype* is that, contingently, things that do either often do both. (FODOR 1998 [37] p. 137)

Le fait de modéliser un lien intentionnel entre deux espaces distincts est déjà problématique en soi, car l'espace conceptuel n'est pas accessible en dehors de ce lien intentionnel. Certes, dans certains modèles, bon nombre de concepts peuvent être accessibles à partir d'autres concepts. Ainsi, si l'on accepte que le concept ÉTALON peut être décrit par une expression comme CHEVAL+MÂLE+ENTIER, on peut espérer définir un mécanisme par lequel le concept ÉTALON hérite son lien intentionnel de ceux des concepts CHEVAL, MÂLE et ENTIER. Toutefois, même dans un schéma définitionnel de ce type, il existe toujours des concepts primitifs qui ne font l'objet d'aucune définition. Comment leur lien intentionnel s'établit-il ? L'expérience produit des percepts, l'exposition au langage permet d'isoler des mots, mais rien, dans l'ontogenèse, ne vient de manière évidente connecter les concepts aux percepts et aux mots qui sont censés les évoquer.

Une solution communément considérée consiste à invoquer l'apprentissage inductif. Ainsi, l'enfant aurait l'occasion de cerner, en généralisant après avoir été exposé à une ou plusieurs occurrences du concept, l'étendue des expériences que le concept est censé reconnaître. Cette solution est incorrecte. Elle présuppose ce qu'elle est censée montrer. Pour reconnaître la première occurrence d'un nouveau concept *P*, l'enfant doit déjà posséder *P* ! Au mieux, le système ne peut que combiner des concepts déjà connus pour produire une formule *G* qui sera équivalente à *P*. Mais un système capable de réaliser une telle opération possède déjà *P* en puissance.

> But notice that learning that [*Px is true iff x is G*] could be learning *P* (learning what *P* means) only for an organism that already understands *G*. For, and this point is critical, *G* in [the preceding] formula is *used*, not mentioned. […] an organism can learn *P* only if it is already able to use at least one predicate that is co–extensive with *P*, viz., *G*.
> (FODOR 1975 [34] p. 80)

De fait, les systèmes d'apprentissage proposés en intelligence artificielle, dans la mesure où il produisent des représentations conceptuelles distinctes des représentations perceptuelles, fonctionnent sur le modèle de l'appariement combinatoire (CF. ANNEXE). L'exposition à des données permet de sélectionner des combinaisons conceptuelles, pas de former des concepts *de novo*. De plus, ces systèmes s'appuient tous sur des concepts primitifs dont le lien avec les données est fourni *a priori*. Il semble donc que le lien intentionnel repose, *in fine*, sur une base innée (FODOR 1975 [34]).

Le problème de la connexion des concepts à leur contenu peut-il ainsi se résoudre par l'hypothèse selon laquelle certains concepts, dits primitifs, disposeraient du privilège d'une intentionnalité innée ? Ce n'est pas certain. Il ne faut pas confondre un concept comme ROUGE, qui est souvent cité pour faire partie du cercle fermé des concepts primitifs, avec la sensation de rouge. Affirmer que le concept ROUGE possède une intentionnalité innée signifie que l'évocation de ce concept est liée, de manière non ambiguë et inamovible, à celle de sa perception. Il devrait être possible de déceler, par des tests psychologiques, la frontière

universelle qui est censée exister entre l'extension perceptuelle d'un tel concept et les percepts qui évoquent d'autres concepts ou qui n'évoquent aucun concept. Rien de tel n'a jamais été montré, et pour aucun concept. Si les sensations connaissent une certaine universalité (CLARK 1993 [20]), il n'existe pas de constante sur les segmentations auxquelles elles donnent lieu (PALMER 1999 [80]). Par ailleurs, invoquer l'innéité des concepts revient à renvoyer la charge d'établissement du lien intentionnel de l'apprentissage à la phylogenèse. Or la sélection naturelle produit des formes localement nécessaires (MONOD 1970 [75]). Il faudrait donc être en mesure d'expliquer que les concepts innés, de par leurs propriétés dans l'espace des concepts, possèdent un lien non arbitraire avec leur contenu, ce qui est loin d'aller de soi (CF. CHAPITRE 8).

Une manière d'échapper aux difficultés liées à l'innéité des concepts primitifs consiste à accepter cette innéité en refusant celle de leur lien intentionnel. L'apprentissage d'un concept pourrait ainsi revenir à une sorte de recrutement d'un concept préexistant pour répondre à un ensemble défini d'expériences. De cette manière, l'important est l'existence d'un mécanisme perceptuel capable de grouper certains percepts, par exemple sous la forme d'un percept moyen, et d'un autre mécanisme capable de connecter ce groupe de perceptions à une représentation de type conceptuel. Ainsi, le concept ROUGE ne se trouverait lié à la perception du rouge que de manière non nécessaire, par le biais d'un mécanisme non inductif de recrutement. Notons que cette solution, si elle permet de résoudre une partie des problèmes évoqués ici, n'en résout pas d'autres, au premier rang desquels celui des liens inter-conceptuels. Nous aurons l'occasion de revenir sur cette question fondamentale de la connexion des concepts à leur contenu (CF. CHAPITRE 8). La solution que nous proposerons consistera à renoncer à l'existence des concepts en tant que représentations mentales permanentes (CF. CHAPITRE 9).

Conclusion

La question fondamentale de l'ancrage des concepts concerne la nature du lien entre les concepts et le monde perçu. Cette question concerne également les théories qui s'intéressent au lien direct que les concepts peuvent avoir avec un monde simplifié considéré comme objectif, puisque la seule connexion qu'un système cognitif peut avoir avec un tel monde s'effectue par le biais de la perception.

La question de l'ancrage se révèle soit triviale, soit extraordinairement ardue. Elle est triviale dans les théories qui ne font pas de différence entre concepts et percepts. Dans ce cas, les concepts sont ancrés de manière naturelle dans la perception. En revanche, dès que les concepts sont considérés comme des représentations distinctes de celles que fournissent nos sens, il devient très difficile d'expliquer comment les concepts se retrouvent liés aux "bons" percepts. Certains auteurs, comme nous l'avons vu, supposent que ce lien est établi par le biais de la communication au sein d'une communauté linguistique, ou du fait d'une disposition cognitive qui nous prédispose à élaborer des représentations conceptuelles portant une information sur le monde. Il n'en reste pas moins que l'amorçage de tels processus reste, dans chaque cas, mystérieux. Comme nous l'avons vu, les "solutions" conventionnelles et informationnelles sont loin de résoudre la question de l'établissement du lien intentionnel. Nous aurons l'occasion de revenir sur cette question fondamentale de l'ancrage (CF. CHAPITRE 8).