UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# UNE STRUCTURE ASSOCIATIVE BIDIRECTIONNELLE D'AUTO-ENCODAGE PERMETTANT L'APPRENTISSAGE ET LA CATÉGORISATION PERCEPTUELS

THÈSE PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR
GYSLAIN GIGUÈRE

NOVEMBRE 2009

# REMERCIEMENTS

Je tiens de tout cœur à remercier plusieurs personnes m'ayant permis d'atteindre mes buts académiques, et m'ayant soutenu tout au long de cette souvent rocambolesque aventure que furent les études supérieures.

Pour débuter, il serait impossible d'oublier le support constant de ma conjointe Stéphanie, qui, dans les bons moments comme lors des tempêtes, a su être présente et patiente en tout temps, et m'offrir son aide, son soutien, et son amour.

Également, je ne pourrais passer sous silence les contributions de plusieurs collaborateurs, mentors et amis, qui m'ont soutenu techniquement, mais aussi professionnellement, moralement et, même dans certains cas, financièrement, tout au long de mes études doctorales : le Dr. Sylvain Chartier (Professeur, École de psychologie, Université d'Ottawa), le Dr. Guy L. Lacroix (Professeur, Département de psychologie, Carleton University), le Dr. Sébastien Hélie (Chercheur postdoctoral, Département de psychologie, University of California – Santa Barbara), ainsi que le Dr. Serge Larochelle (Directeur, Département de psychologie, Université de Montréal).

Finalement, je désire aussi remercier mes directeurs de recherche, les professeurs Jean-Marc Lina (Département de génie électrique, École de Technologie Supérieure) et Robert Proulx (Vice-recteur à la recherche et à la création, UQÀM), dont les connaissances, les compétences, ainsi que le soutien théorique et technique m'ont permis de concevoir ce projet et de le mener à terme.

Cette thèse a été rendue possible grâce au support financier du Conseil de Recherches en Sciences Naturelles et en Génie (CRSNG) et du Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT).

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

# RÉSUMÉ

Les humains sont continuellement exposés à des stimulations pour lesquelles leur système perceptivo-cognitif doit créer des représentations mnésiques. Tout en créant un code interne de composantes, ce système doit être en mesure de reconnaître, d'identifier, et de discriminer ces objets lors de prochaines occurrences. Ce processus s'effectue par la création et la mise à jour d'une mémoire épisodique d'exemplaires à dimensionnalité réduite. De plus, le système cognitif doit regrouper les objets similaires en catégories, tout en adaptant le contenu de la mémoire suite à l'ajout d'informations produit par la rencontre de nouveaux objets. Ces processus de niveau « objet » et « catégorie » s'effectuent de façon séparée, par le biais de deux mémoires.

Jusqu'à maintenant, aucun modèle formel satisfaisant n'était en mesure de rendre compte de cette variété de comportements humains sans sacrifier la simplicité et l'élégance du système initial pour simuler l'un d'eux. Le modèle FEBAM (pour *Feature-Extracting Bidirectional Associative Memory*) a été créé dans le but de répondre à cette incapacité de beaucoup de modèles existants à effectuer des tâches cognitives et perceptuelles à l'aide d'un codage interne créé de façon autonome, comme le font les humains. Basé sur une architecture neuronale associative bidirectionnelle, FEBAM peut reproduire les comportements d'autres réseaux de neurones artificiels dont les processus dynamiques sont basés sur l'extraction de composantes, la création de bassins d'attracteurs, ou encore le partitionnement de données (« clustering »), et ce, en utilisant une seule architecture, règle de transmission et procédure d'apprentissage.

Dans la présente thèse, il sera montré qu'avec un nombre minimal de principes définitoires, le modèle pourra effectuer des tâches telles que la création autonome d'un code interne de composantes, le développement autonome d'une mémoire d'exemplaires parfaits, ainsi que l'identification et la catégorisation autonomes. Il sera aussi montré, grâce à la proposition d'un mécanisme itératif de croissance de l'architecture, que les catégories créées par le réseau peuvent être réorganisées suite à la présentation de nouvelles informations perceptuelles au système. On montrera également que FEBAM préserve les capacités d'une mémoire autoassociative récurrente (dont il est inspiré), tout en améliorant certains des comportements de cette dernière.

Le modèle FEBAM sera également étendu au cas supervisé. Dans ce cas, le modèle FEBAM-RA (RA pour *Response Association*), grâce à un module supplémentaire, associera les représentations internes des stimuli à leur identité ou à leur appartenance catégorielle prédéfinies. Cette extension se fera sans avoir à ajouter des principes définitoires: ainsi, on utilisera ici la même règle d'apprentissage, la même règle de transmission, et une généralisation de l'architecture de FEBAM. Grâce à cet ajout, le modèle sera en mesure de reproduire de façon qualitative l'effet de la pré-exposition perceptuelle sur la rapidité de l'apprentissage identificatif supervisé, ainsi que l'effet de difficulté de la tâche lorsque l'on

compare l'identification et la catégorisation supervisées (dans une situation de tâches simultanées).

La contribution principale de cette thèse repose donc dans la parcimonie des principes utilisés. En effet, grâce à un nombre minimal de postulats définitoires, on modélisera donc des processus de traitement d'objets et de catégories, et ce, de façon autonome ou supervisée. Ce projet de recherche constituant la première étape de développement de l'approche FEBAM, quelques améliorations à l'approche de base seront proposées.

Mots-clés : modélisation cognitive, réseaux de neurones artificiels, extraction de composantes, catégorisation, identification.

# INTRODUCTION

Les humains sont continuellement exposés à des situations durant lesquelles ils doivent différencier, reconnaître, identifier, et regrouper des patrons d'entrée perceptuels (tels que les objets présentés au champ visuel). Leur système perceptivo-cognitif effectue ces opérations dans le but de produire des réponses et actions appropriées, en fonction de l'identité et des propriétés des stimuli rencontrés. Pour effectuer ces opérations, le système doit créer et enrichir des représentations mnésiques adaptées à différents environnements nécessitant une gamme de comportements réactifs différents (dépendant du contexte). Ce processus général, connu sous le nom d'apprentissage perceptuel, consiste principalement en l'abstraction implicite d'information précédemment non-disponible au système, dans le but de produire des changements semi-permanents au niveau des structures mnésiques (Gibson et Gibson, 1955). La plupart des processus d'apprentissage perceptuel peuvent être réalisés de façon autonome, par l'abstraction associationiste des structures statistiques environnementales (comme le font certains réseaux de neurones : Goldstone, 1998; Hall, 1991).

*Traitement perceptuel d'objets*

Pour représenter les stimuli entrants en mémoire, le système cognitif humain doit définir et utiliser un ensemble de composantes (Garner, 1974). Ces composantes seront utiles lors de situations futures, lorsqu'une décision basée sur une stimulation perceptuelle est nécessaire. Dans les modèles symboliques (Newell et Simon, 1972), l'identité et la composition de ces caractéristiques perceptuelles sont prédéterminées par le modélisateur. Cependant, dans la vraie vie, les systèmes cognitifs humains doivent créer leur propre système ou vocabulaire de composantes, de façon autonome, en utilisant les propriétés statistiques de l'environnement (Harnad, 1990; Schyns et Rodet, 1997). L'empreinte perceptuelle laissée par les composantes acquises implique que ces dernières doivent d'abord être extraites des patrons entrants. Cette extraction devrait mener à la création d'un système définitoire d'objets perceptuels, un ensemble de « blocs de construction » perceptuels (Schyns, Goldstone et Thibaut, 1998). Les

composantes peuvent ensuite réutilisées pour traiter de nouveaux patrons d'entrée rencontrés. Le système perceptivo-cognitif humain a donc avantage à créer un système de composantes qui inclut des propriétés statistiques menant à une reconnaissance, une identification et une différentiation rapides (Goldstone, 1998).

*Catégorisation perceptuelle*

L'invariance et la pertinence sont des qualités qui tendent à accélérer les processus cognitifs (et plus précisément, la récupération de traces mnésiques), en réduisant la quantité d'information devant être prise en compte dans une situation donnée. Ce principe est celui de l'économie cognitive (Goldstone et Kersten, 2003). Pour effectuer ces économies, le système doit être enclin à effectuer de la réduction d'information, en regroupant des stimulations perceptuelles similaires et en les représentant par des entités abstraites, créées en fonction des caractéristiques communes des objets. Ce processus est celui de la catégorisation perceptuelle.

Historiquement, les psychologues cognitifs se sont peu entendus à propos de la nature des représentations utilisées par le système perceptivo-cognitif pour la catégorisation et la classification. Certains croient que le système utilise des abstractions génériques (*e.g.*, des prototypes; par exemple : Reed, 1972; Smith et Minda, 2002), alors que d'autres affirment qu'il utilise des stimulations perceptuelles spécifiques (*i.e.* des exemplaires; par exemple : Medin et Shaffer, 1978; Nosofsky, 1984, 1986). Les auteurs qui supportent l'idée des prototypes sont d'avis que les catégories possèdent un statut représentationnel spécial dans le système. Pour eux, chaque exemplaire du monde extérieur est lié à une représentation générique sommaire, pouvant être récupérée indépendamment, et pour laquelle la trace mnésique est plus forte et plus durable que pour les exemplaires associés (Posner et Keele, 1968, 1970). Ces représentations aident le système à décider rapidement de l'appartenance catégorielle, et sont enrichies (ou réorganisées) selon les nouvelles informations perceptuelles disponibles (Murphy, 2002). Pour que le système réussisse ceci, les catégories doivent être faciles à différencier; c'est pourquoi le monde naturel est composé de catégories cohésives, à l'intérieur desquelles les exemplaires associés sont similaires les une aux autres (Rosch, 1973).

*Modélisation cognitive de systèmes de caractéristiques*

La grande majorité des cognitivistes croit que le système cognitif utilise un ensemble de caractéristiques (ou composantes) pour effectuer les tâches perceptuelles. Ainsi, les systèmes de composantes sont un principe généralement accepté en cognition. La plupart des chercheurs et théoriciens voient les caractéristiques et dimensions d'objets comme des définitions symboliques (instanciations *digitales*) et prédéterminées. Depuis une dizaine d'années, certains auteurs ont tenté de convaincre la communauté scientifique que ces systèmes devaient être vus de façon perceptuelle, analogique et surtout, flexible. Les composantes d'objets seraient définies d'une façon abstraite et iconique. Elles seraient créées de façon implicite, principalement selon des processus associatifs ascendants, basés sur une stratégie de réduction de l'information (Harnad, 1990; Schyns et al., 1998; Schyns et Murphy, 1994; Schyns et Rodet, 1997).

Selon plusieurs auteurs, cet ensemble d'opérations peut être effectué de façon autonome (*i.e.* de façon non-supervisée : Hall, 1991), et mène à la création d'un système de « blocs de construction » perceptuels. Ce système, pour être utile et économique, doit mener vers la différentiation rapide des objets et des catégories (Goldstone, 1998). Aussi, il devrait être réutilisable, pour que de nouveaux objets puissent être représentés à l'aide du code existant. Au niveau théorique, l'idée d'un système qui produit son propre système de composantes perceptuelles, composé de représentations « atomiques », fait lentement son chemin. Malheureusement, peu d'efforts ont été faits pour modéliser ce type de processus, spécialement dans le domaine de la catégorisation perceptuelle (Schyns et al., 1998).

*But de la thèse*

Le but principal de cette thèse est de proposer un nouveau modèle (ainsi qu'une extension de ce modèle) permettant de reproduire des processus d'apprentissage et de catégorisation perceptuels. Plus précisément, on voudra modéliser la création d'un vocabulaire de composantes iconiques, la réduction dimensionnelle (économie cognitive), la différentiation d'objets, la catégorisation autonome, ainsi que l'identification et la catégorisation supervisés. Toutes ces tâches perceptivo-cognitives devront être modélisés à partir des mêmes principes définitoires. On visera donc ici principalement la parcimonie et la

simplicité. Ainsi, tous les processus modélisés utiliseront le même type d'architecture neuronale, la même règle d'apprentissage, la même règle de transmission, et ce, qu'ils soient réalisés de façon autonome ou supervisée.

Il est à noter qu'en ce premier stade de développement de l'approche, on ne vise pas encore l'adéquation de données empiriques (*data fitting*). On voudra plutôt s'assurer de reproduire des tâches spécifiques, ainsi que des effets trouvés dans la littérature, mais au niveau qualitatif.

*Présentation de la thèse*

La présente thèse sera divisée comme suit. Le premier chapitre présentera diverses problématiques liées à l'apprentissage et à la catégorisation perceptuels. On y explorera plus spécifiquement les théories de différentiation d'objets et d'apprentissage catégoriel, en plus de traiter de différents effets montrant l'existence d'un processus d'extraction de composantes perceptuelles. Dans le second chapitre, on définira des caractéristiques désirables pour un modèle perceptivo-cognitif. Ces hypothèses de travail seront principalement inspirées de théories et modèles d'apprentissage et de catégorisation perceptuels explorés au Chapitre 1. Ce chapitre explorera les différentes classes de réseaux de neurones existantes, tentant de déterminer si l'une ou l'autre de ces classes peut entièrement répondre aux problèmes posés. Le Chapitre 3 présentera les caractéristiques de base du réseau FEBAM (Feature-Extracting Bidirectional Associative Memory), un auto-encodeur auto-supervisé à connexions asymétriques permettant l'extraction de caractéristiques, le développement d'une mémoire d'exemplaires, ainsi que la catégorisation autonome. Cette proposition devrait répondre à toutes les exigences posées en ce qui a trait à l'apprentissage des objets et des catégories.

Au Chapitre 4, on explorera les capacités du réseau FEBAM en ce qui a trait aux tâches perceptivo-cognitives désirées. Aussi, on comparera FEBAM au modèle duquel il émane, pour s'assurer que le modèle conserve les caractéristiques désirées d'une mémoire autoassociative récurrente, malgré l'ajout d'une couche d'unités. Au Chapitre 5, une extension de FEBAM, FEBAM-RA (RA pour *Response Association*), sera présentée. Ce modèle devrait être en mesure de répondre aux exigences de tâches avec rétroaction externe,

mais sans devoir recourir à la rétropropagation. Des simulations montrant les caractéristiques et comportement du réseau seront présentées. Finalement, le Chapitre 6, en plus de résumer la thèse, présentera quelques perspectives de recherche possible pour améliorer la performance et l'adéquation du réseau en ce qui a trait à la modélisation de tâches expérimentales psychologiques.

Pour terminer, notons que plusieurs des simulations présentées dans cette thèse ont été publiées dans des articles d'actes de conférence, tels que ceux de l'International Conference on Cognitive Modeling (Giguère, Chartier, Proulx, et Lina, 2007a), de la Cognitive Science Society (Giguère, Chartier, Proulx, et Lina, 2007b), ainsi que de l'International Joint Conference on Neural Networks (Chartier, Giguère, Renaud, Lina, et Proulx, 2008).

# CHAPITRE I

## CONSIDÉRATIONS PERCEPTIVO-COGNITIVES

Selon la définition originale de Gibson et Gibson (1955), l'apprentissage au niveau perceptuel consiste en l'abstraction d'information environnementale invariante (i.e., commune aux différentes entrées perceptuelles), préalablement non-disponible au système cognitif (d'où le terme « apprentissage »). Les mécanismes principaux de ce type d'apprentissage seraient : 1) la détection des caractéristiques composant les objets; et 2) l'abstraction de propriétés plus globales et générales, telles les relations entre les objets (Gibson, 1969). Ces processus globaux émanent du fait que tout système perceptivo-cognitif inclut une mémoire où chaque entrée (ou sa représentation) est reliée d'une certaine façon aux autres entrées déjà traitées par le système (Gibson, 1969; Shepard, 1958). Solley et Murphy (1960; dans Gibson, 1963) expliquent cette globalité inhérente : selon eux, à mesure que des échantillons perceptuels (ou « sensoriels ») sont accumulés dans la mémoire, des points de référence cognitifs sont développés. Par la suite, les nouvelles entrées doivent en quelque sorte être interprétées en fonction des points de référence existants. Le thème de la globalité est aussi soulevé par Lashley (1950, dans Gibson, 1969), qui considère que le processus d'apprentissage perceptuel en est un d'harmonisation entre les différents éléments composant le système complexe perceptif (et donc une harmonisation globale). Ces éléments se veulent selon lui des combinaisons de patrons (stimuli) et d'unités (composantes) qui répondent de façon différente suite à l'apprentissage.

De nos jours, il est généralement accepté que la majorité des phénomènes d'apprentissage perceptuel peuvent être réalisés de façon autonome (non-supervisée) par l'abstraction des structures statistiques inhérentes à cet environnement (tout comme le font les réseaux de neurones non-supervisés; Hall, 1991). Cet apprentissage peut être réalisé suite à la seule exposition au stimulus (Goldstone, 1998). Solley et Murphy (1960; dans Gibson,

1963) soulignent la nécessité de cette autonomie. Selon eux, les humains doivent posséder une motivation intrinsèque à « mettre de l'ordre » dans les nombreuses stimulations perceptuelles provenant de l'environnement, sans quoi la perception ne possède aucune utilité fonctionnelle.

On considère habituellement les résultats de l'apprentissage perceptuel comme « semi-permanents ». Ainsi, les résultats induits par le processus sont présents à long terme dans le système (concept de stabilité : Gibson, 1963; Grossberg, 1987), mais peuvent être modifiés pour améliorer la capacité du système à répondre aux changements dans son environnement perceptuel (concept de plasticité : Grossberg, 1987; Ittelson, 1962). Ittelson (1962) souligne d'ailleurs, au sujet de la plasticité, que la perception implique une réévaluation constante de l'environnement, et constitue un processus créatif au cours duquel chaque individu construit pour lui-même son propre « monde d'expériences ». Gibson (1963) relève l'utilité de ce principe en parlant d'un système cognitif voué à structurer et restructurer son environnement de façon active.

Selon Gibson (1963), l'apprentissage perceptuel mène donc à la découverte d'une structure, d'un ordre, d'une stabilité environnementale. Bruner, Wallach et Galanter (1959) indiquent que la force de l'apprentissage perceptuel repose sur la technique utilisée pour déterminer l'identité et la pertinence des composantes de l'entrée (par la découverte de régularités environnementales). Barlow (1961, 1989) croit que cette dite technique doit nécessairement reposer sur la réduction de la redondance, ou la compression, mécanisme selon lui essentiel de l'encodage des messages perceptuels. L'idée de retirer la redondance des entrées perceptuelles et d'ainsi réduire la taille des représentations lui semble naturelle, puisque la discrimination entre les différentes stimulations perceptuelles est beaucoup plus facile si l'information sensorielle est traitée de façon comprimée.

Goldstone (1998) rapporte que l'apprentissage perceptuel peut être divisé en deux classes principales, soit le niveau objet et le niveau catégorie. De ces classes, plusieurs sous-processus peuvent être explorés, dont : l'empreinte originale de stimuli et de caractéristiques perceptuelles (*i.e.* leur mémorisation complète ou « comprimée »), la différentiation entre ces stimuli encodés, ainsi que la création et différentiation de catégories. Ces différents thèmes seront explorés au cours du présent chapitre.

## 1.1 Niveau objet

D'après les créateurs de certains modèles exemplaristes d'automatisation des apprentissages (Logan, 1988) et de catégorisation (Nosofsky, 1986), l'une des tâches effectuées par le système cognitif au niveau perceptuel est l'encodage des exemplaires présentés (ou la création « d'empreintes » de stimuli : Gibson, 1969). La vision de ces auteurs (Nosofsky en particulier) mène à une thèse exemplariste forte, où chaque item est mémorisé de façon complète (principe d'encodage obligatoire : Nosofsky et Palmeri, 1997). Plusieurs cognitivistes (dont Valentin, Abdi et Edelman, 1998) considèrent plutôt que cet encodage se fait de façon imparfaite; certains affirment d'ailleurs que la mémoire se construit de façon optimale au niveau perceptuel, en ne conservant que l'information permettant un traitement subséquent adapté[1].

L'empreinte de ces caractéristiques dans le système cognitif implique qu'elles doivent tout d'abord être extraites du stimulus entrant, dans le but de créer un système de « définitions perceptuelles » d'objets. Cet assortiment de caractéristiques, tel un ensemble de « blocs de construction » (*building blocks* : Schyns et al., 1998), peut ensuite être réutilisé pour traiter un nouvel objet dans un contexte différent. Le système cognitif a avantage à créer un ensemble efficace de représentations dimensionnelles, en y incluant les caractéristiques statistiques de l'environnement menant à une différentiation rapide des exemplaires et des catégories (Goldstone, 1998). Tel qu'il sera montré, l'extraction de caractéristiques constitue un processus jugé de plus en plus crucial en psychologie cognitive, mais est en fait rarement pris en compte dans les modèles cognitifs en général, et dans les modèles de catégorisation en particulier (Harnad, 1990; Schyns et al., 1998).

En plus d'encoder et de mémoriser les stimuli selon un système interne de caractéristiques, le système cognitif doit ensuite pouvoir différencier les stimuli. Le fait que les objets originalement confondus deviennent plus faciles à discriminer suite à la pratique (*i.e.* que leur caractère différentiel s'accroisse) est maintenant généralement accepté par les

[1] Malgré que l'utilisation des exemplaires comme base des catégories s'avère passablement polémique (tel que détaillé dans la section suivante), beaucoup de cognitivistes s'entendent pour accorder au système cognitif un rôle de mémoire épisodique d'une quelconque représentation d'exemplaires, même s'ils ne croient pas que le système utilise la totalité des exemplaires mémorisés lors de décisions catégorielles (Smith, 2005).

théoriciens de l'apprentissage perceptuel (Goldstone, 1998). En fait, selon Gibson et Gibson (1955; Gibson, 1969), la différentiation constitue le but principal de l'apprentissage perceptuel. La pratique permet de préciser la représentation de chaque stimulus. Cette précision apprise réduit la confusion, et par le fait même, réduit graduellement le pouvoir d'attraction des représentations d'objets dans le système. En augmentant le nombre de représentations, on réduit ce pouvoir d'attraction, ce niveau de généralisation aux stimuli similaires. Avec l'expérience, le système, qui à l'origine ne contient que peu d'information de nature à confondre tous les objets, pourra graduellement inclure des traces mnésiques distinctes pour ces derniers.

Plusieurs preuves appuient ce principe de différentiation. Par exemple, il a été montré à maintes reprises que la simple exposition rapide à un stimulus améliore la performance à une tâche d'identification (Schacter, 1994). Également, selon la méthode établie par Gibson et Walk (1956), une préexposition à un ensemble de stimuli (une tâche, notons-le, sans but explicite) permet d'améliorer les performances à une tâche subséquente de catégorisation utilisant ces mêmes stimuli. Cet effet a d'ailleurs été montré à de nombreuses reprises, avec des animaux devant différencier deux classes de cercles et de triangles (Gibson et Walk, 1956), et aussi avec des humains, qui devaient par exemple discriminer parmi un ensemble de « gribouillis » (*squiggles*) originalement impossibles à distinguer (Goldstone, 1998). Ces résultats montrent la « motivation intrinsèque » du système à traiter les objets de façon perceptuelle. Ils soutiennent aussi l'hypothèse selon laquelle la prédifférentiation est le but principal du système perceptivo-cognitif[2].

La différentiation de stimuli est liée à trois tâches classiques en cognition, soit la reconnaissance, la discrimination, et l'identification absolue (Nosofsky, 1984; 1986). La reconnaissance nécessite de pouvoir déterminer si un exemplaire présenté éveille une trace mnésique dans un temps raisonnable, ou non. Dans le cas positif, le système doit indiquer qu'il a vu l'exemplaire, mais n'a pas à l'identifier spécifiquement. L'identification absolue est intimement liée à la reconnaissance : en plus de récupérer la trace mnésique, le système doit également recouvrer l'étiquette unique associée à l'objet suite à un apprentissage

[2] Hall (1991) rapporte que malgré que toutes les réplications du paradigme original Gibson-Walk n'aient pas produit des résultats aussi forts qu'à l'origine, cet effet est généralement accepté comme valide.

précédent. Enfin, dans une tâche de discrimination, on présente deux exemplaires au système, de façon séquentielle ou simultanée, et le système doit déterminer si les exemplaires sont différents ou identiques. Si la présentation force la récupération de deux représentations distinctes, c'est que les objets sont jugés différents. Sinon, c'est que le même objet a été présenté deux fois.

Ainsi, la présentation d'un exemplaire au système mènera à l'activation d'une trace mnésique existante et à son possible enrichissement (par création de nouvelles caractéristiques par exemple), ou encore à la création d'une nouvelle trace (ou d'une nouvelle caractéristique) pour accommoder l'objet perçu. Selon Gibson et Gibson (1955), le principe de différentiation n'exclut pas celui de l'enrichissement. Ainsi, en tentant de mieux discriminer les stimuli, le système produira des traces mnésiques de plus en plus informatives.

## 1.2 Niveau catégorie

Le système perceptivo-cognitif, s'il peut encoder des représentations individuelles de patrons d'entrée, tend également à créer des représentations plus abstraites, englobant les caractéristiques communes à plusieurs stimulations perceptuelles distinctes. Ce processus concerne donc l'empreinte, la différentiation et l'enrichissement de ce que l'on nomme les catégories.

La catégorisation, telle qu'accomplie par un système perceptivo-cognitif, peut se définir comme la mise en commun psychologique d'entités perceptuelles distinctes (tels les objets visuels), dans le but de faciliter la prise de décision et la réponse appropriée face aux différents objets rencontrés. Il s'agit d'une opération cognitive cruciale et répandue, dont l'omniprésence émane de la propension naturelle qu'ont les humains à interpréter leur environnement, plutôt que de se limiter à le percevoir passivement (Wittgenstein, 1953), *i.e.* à voir « toute chose » comme « quelque chose » (*to perceive a thing as something* : Goldstone et Kersten, 2003). Tout humain[3], adulte ou enfant (Bomba et Siqueland, 1983), se sert

[3] Ainsi que de nombreux animaux tels les singes macaques (Smith, Redford, et Haas, 2008), les pigeons (Aydin et Pearce, 1994) et même les abeilles (Benard, Stach, et Giurfa, 2006)!

d'habiletés dites conceptuelles ou catégorielles[4] pour différencier des groupes de stimulations perceptuelles qui, par exemple, frappent sa rétine. L'ubiquité de ce processus pousse même Harnad (2005; Murphy, 2002) à affirmer que tout acte cognitif en est un de catégorisation (« to cognize is to categorize : cognition is categorization »; Harnad, 2005, p.1).

En général, les cognitivistes s'entendent pour déterminer un statut représentationnel spécial à la catégorie. Outre les exemplaristes, la plupart des scientifiques croient que tout ensemble de stimulations perceptuelles communes est associé à une représentation mentale plus sommaire (règle perceptuelle ou verbale, prototype, zone définie par une frontière décisionnelle), pouvant être récupérée de façon indépendante, et dont la trace mnésique est plus forte et durable que pour l'ensemble des exemplaires (*e.g.*, Posner et Keele, 1970; Smith et Minda, 2002). Les représentations ainsi créées doivent aider le système à déterminer précisément et rapidement l'appartenance catégorielle d'un stimulus. Pour ce faire, les catégories doivent être différenciables facilement. C'est pourquoi le monde naturel serait composé de catégories cohésives, pour lesquelles les exemplaires représentés sont hautement similaires entre eux, mais peu similaires aux membres d'autres catégories (Rosch, 1973).

Selon Goldstone et Kersten (2003; voir aussi Komatsu, 1992; Murphy, 2002; Smith et Medin, 1981), la catégorisation est utile à de nombreux égards. Notre paysage catégoriel mental nous sert de filtre cohérent, organisateur, à travers lequel sont interprétées les stimulations du monde extérieur[5]. Ainsi, les représentations mentales de notre environnement ne sont pas « pures » (au sens où la mémoire à long terme serait pleinement iconique ou photographique; Lacroix, Giguère et Larochelle, 2005), mais plutôt mentalement « restructurées » dans un but ultime, soit la prise en compte et la mémorisation de l'information diagnostique (*i.e.* utile à la différentiation psychologique et au processus décisionnel).

L'une des fonctions les plus importantes de cette restructuration est l'économie cognitive (Goldstone et Kersten, 2003; Komatsu, 1992), qui peut être illustrée de diverses

[4] Précédemment, les termes « catégorie » et « concept » pouvaient présenter des significations distinctes (la catégorie étant l'ensemble d'objets dans le monde réel, et le concept, la représentation mentale), depuis une dizaine d'années, les deux sont utilisés de façon interchangeables par la plupart des auteurs (Murphy, 2002).

[5] Murphy (2002) met l'accent sur la cohérence et la cohésion en affirmant que « les concepts sont la colle qui unifie notre monde mental » (p.1).

manières. D'un point de vue brut d'espace de stockage cognitif, des économies substantielles sont possibles si l'on encode une seule représentation abstraite[6]. Celle-ci contiendrait les informations diagnostiques et communes aux différents membres de la catégorie, plutôt que toute l'information spécifique définissant chaque exemplaire rencontré. Ainsi, en plus de réduire le nombre de représentations nécessaires, la taille de chaque représentation résultante est aussi diminuée, laissant de côté l'information jugée non-pertinente par le système. En termes de la théorie de l'information, la catégorisation permet donc d'optimiser le code représentationnel, en réduisant le nombre de bits définitoires (Anderson, 1991).

En plus de préserver la mémoire, la catégorisation permet de réduire la nécessité d'apprendre grâce au principe de généralisation. En effet, si une nouvelle stimulation est très similaire à l'une des représentations catégorielle développées, un réapprentissage complet est alors inutile. Cependant, si cette stimulation est quelque peu semblable à une représentation connue, une mise à jour ou une légère réorganisation du paysage catégoriel (Murphy, 2002) peut suffire. Si l'on adopte un point de vue plus associationiste (de type contingences entre un stimulus et une réponse), la catégorisation permet de réduire le nombre d'associations à apprendre entre les stimulations du monde extérieur et les réponses adéquates associées. La compression des représentations permet également d'accélérer la récupération de la trace mnésique associée à une stimulation perceptuelle, et permet donc de produire une réponse appropriée plus rapidement.

Il existe un grand nombre de théories et de modèles de catégorisation. Le principal point de distinction (et de discorde) entre les théories concerne le type de représentation mentale postulée. Les prochaines sections présenteront les principales approches proposées au cours des dernières décennies, tout en mettant en relief les propriétés de chacune, leurs avantages, les critiques auxquelles elles font face, ainsi que les éléments désirables et récupérables qui sont rattachés à chacune des approches.

[6] C'est à dire suivant un processus d'abstraction, de réduction.

### 1.2.1 L'approche par règles/définitions

Originalement, les premiers énoncés sur la catégorisation furent proposés par Platon et Aristote, qui définirent la nécessité de grouper les objets liés selon leurs caractéristiques verbalisables communes en ces entités mentales, discrètes et cohérentes que sont les catégories (aussi nommés « concepts »). Les philosophes ont argumenté à propos de ce sujet durant plusieurs siècles, produisant des discussions teintées du conflit entre nativisme et empirisme. Il fallut l'avènement de la psychologie scientifique (à la fin du 19e siècle) pour enfin penser à déterminer empiriquement les bases et les propriétés des processus de formation et d'utilisation des catégories.

D'un point de vue scientifique, l'un des premiers auteurs à étudier la catégorisation de façon empirique fut Hull (1920, dans Murphy, 2002), qui fit apprendre à des participants des catégories artificielles de façon inductive[7]. Les stimuli utilisés par Hull étaient des caractères chinois. Chaque membre complexe d'une catégorie contenait une racine graphémique simple dénotant son appartenance catégorielle. Si le membre contenait la racine, il était membre de la catégorie; sinon, il n'en était pas membre (principe « tout-ou-rien.»). Hull, suite au succès des participants, conclut que ces derniers avaient abstrait une règle logique permettant de déterminer l'appartenance catégorielle d'après la racine perçue.

Bruner, Goodnow et Austin (1956) s'inspirèrent de ces résultats et des principes établis de la logique du premier ordre pour proposer la vision dite « classique » (Smith et Medin, 1981) de la catégorisation. Selon cette approche, chaque catégorie du monde extérieur est abstraitement représentée dans le système cognitif par une définition ou règle verbale, basée sur un ensemble de caractéristiques individuellement nécessaires et conjointement suffisantes. Le caractère de nécessité signifie qu'un item doit posséder la totalité d'un ensemble de caractéristiques définitoires pour appartenir à une catégorie, alors que le principe de suffisance implique que lorsqu'un item contient toutes les caractéristiques définitoires sélectionnées pour une catégorie, il en est obligatoirement un membre. En termes logiques,

[7] L'induction catégorielle, qui est encore de nos jours le paradigme expérimental de choix en catégorisation, implique la création d'ensembles de stimuli appartenant à deux (ou plusieurs) catégories artificielles. Aucune information préalable sur la composition des catégories n'est fournie aux participants. Lors de chaque essai, on présente un stimulus au participant, qui doit décider de l'appartenance catégorielle. Ce dernier reçoit ensuite une rétroaction corrective, et doit se servir de celle-ci pour découvrir la composition des catégories.

les caractéristiques sont des prémisses, et la catégorie, une conclusion.

La vision classique constitue, sous certains points de vue, une approche optimale de la catégorisation, puisqu'elle se base sur une logique précise de type « tout-ou-rien », constituant ainsi un modèle inférentiel puissant. Cela n'a pas empêché cette approche d'être rapidement déclarée invraisemblable (Smith et Medin, 1981; Medin, 1989), pour des raisons théoriques telles que : 1) la quasi-impossibilité de trouver des ensembles de caractéristiques individuellement nécessaires et conjointement suffisantes pour la plupart des catégories (Wittgenstein, 1953, dans Laurence et Margolis, 1999); 2) la présence courante de définitions disjonctives dans le langage naturel (Rosch, Mervis, Gray, Johnson et Boyes-Braem, 1976). Au niveau empirique, la découverte de l'effet de typicité (Rosch, 1975; Rosch et Mervis, 1975) fut dévastateur pour l'approche classique. En effet, selon cette vision, tous les exemplaires d'une catégorie possédaient un statut équivalent. Toutefois, plusieurs auteurs ont montré que certains exemplaires constituaient de meilleurs représentants de leurs catégories respectives. Ces exemplaires dits « typiques » différaient entre autres en ce que les participants pouvaient les identifier et les classer plus rapidement (Murphy et Brownell, 1985; Rips, Shoben et Smith, 1973), les apprenaient de façon prioritaire (Rosch, Simpson, et Miller, 1976) et les nommaient en premier lorsqu'on leur demandait de lister les membres d'une catégorie (Mervin, Catlin, et Rosch, 1976). Les exemplaires typiques étaient donc considérés comme plus centraux, plus « rapprochés » de la représentation catégorielle.

### 1.2.2 Les approches à base de similarité

Le principe de typicité mit en valeur la notion voulant que l'on puisse postuler une mesure de ressemblance psychologique entre différentes représentations mentales. De façon convergente, des auteurs comme Shepard (1957, 1958) ont à l'époque proposé l'idée de la cognition humaine comme un espace topologique basé sur une métrique de distance psychologique. Les avancées dans le domaine de l'analyse multidimensionnelle (Torgerson, 1952; voir aussi Shepard, 1962) ont porté Shepard à proposer que les objets étaient représentés en mémoire tels des points ou positions dans un hyperespace en $n$ dimensions, et que la similarité psychologique pouvait être quantifiée grâce à des postulats de transformation exponentielle (Shepard, 1987) et de distance basée sur les principes de Minkowski (voir

Nosofsky, 1986). Plusieurs expériences perceptuelles ont confirmé ces postulats, montrant par exemple que les catégories de couleurs étaient organisées de façon circulaire au niveau psychologique (Shepard, 1962). Selon Edelman et Intrator (1997), l'introduction du concept d'espace multidimensionnel psychologique tel que présenté par Shepard est très important, puisqu'il retire aux nouveaux stimuli leur « statut distinct ». Ces derniers sont également des points dans l'espace, et leur position n'a qu'à être caractérisée à l'aide de leur position parmi les points (stimuli) déjà connus.

Faces aux lacunes de l'approche classique, certains cognitivistes ont proposé de combiner cette nouvelle vision basée sur les métriques et distances psychologiques aux récents principes symboliques fondateurs de Newell et Simon (1972) (inspirés de la thèse Church-Turing[8]). Ils ont ainsi proposé diverses approches computationnelles dites « symboliques ». Selon la théorie des symboles physiques, tout système « intelligent » se définit comme une machine permettant le traitement algorithmique d'informations représentées par une concaténation de patrons structurés (ou « symboles ») sous forme « numérique » (en anglais : *digital*). Cette information est alors traitée par le système cognitif[9] selon un ensemble de règles ou d'algorithmes pour produire des connaissances supplémentaires pouvant être mémorisées et réutilisées.

La fonctionnalité des systèmes basés sur cette approche repose sur une définition des connaissances sous forme de listes de symboles pouvant prendre une valeur entière ou continue. En catégorisation, chaque symbole représente une caractéristique ou dimension (simple ou complexe) du problème à résoudre. Lorsque ce problème implique le traitement d'un objet perceptuel, les valeurs de dimensions (qui à ce point-ci, se doivent d'être verbalisables) peuvent représenter, par exemple, la présence ou absence d'une composante, le statut d'une caractéristique pouvant présenter plusieurs valeurs différentes, ou encore une mesure objective discrète ou continue (Garner, 1974)[10].

---

[8] Selon la thèse Church-Turing, tout problème logique ou mathématique peut être résolu par un système formel suivant des « procédures efficaces » (ou algorithmes), c'est-à-dire une description complète et non-ambiguë d'un ensemble fini d'opérations pouvant être réalisées de façon précise selon une procédure strictement mécanique.

[9] Compte tenu de la polémique entourant les multilples définitions de l'intelligence, le terme « système intelligent » est ici remplacé par "système cognitif", un terme plus neutre.

[10] Selon Garner (1974), ces dimensions, au niveau symbolique, peuvent être extraites par le système cognitif de façon indépendante (dimensions séparables) ou simultanée et indifférenciée (dimensions intégrales).

**Tableau 1.1** Structure catégorielle symbolique de ressemblance familiale[11]

| Catégorie A | | | | Catégorie B | | | |
|---|---|---|---|---|---|---|---|
| Dim. 1 | Dim. 2 | Dim. 3 | Dim. 4 | Dim. 1 | Dim. 2 | Dim. 3 | Dim. 4 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

Les approches symboliques de catégorisation sont généralement basées sur des mesures de similarité globales entre des listes de symboles prédéfinis. Le Tableau 1.1 présente une structure catégorielle symbolique. On voit que la catégorie A contient principalement des valeurs de 0, alors que la catégorie B contient principalement des valeurs de 1. Les valeurs 0 et 1 représentent, pour chaque dimension (couleur ou longueur, par exemple) utilisée dans la construction des stimuli expérimentaux, deux implémentations visuelles discrètes différentes (par exemple : rouge vs. vert, long vs. court, etc.).

#### 1.2.2.1 L'approche prototypiste

La première approche basée sur la similarité fut la vision prototypiste (Reed, 1972; Rosch, dans Rosch et Lloyd, 1978), qui base les décisions catégorielles sur un calcul de similarité entre un nouvel exemplaire à classer et une tendance centrale développée progressivement lors de l'apprentissage. Le modèle original veut qu'en termes symboliques, une entité X sera catégorisée comme membre d'une catégorie Y si et seulement si la somme pondérée du nombre de valeurs « symboliques » communes à X et au prototype de Y dépasse un seuil critique (Smith et Medin, 1981).

Le principe de base est que par apprentissage inductif, les participants à une tâche de catégorisation en viendront à développer deux représentations maximalement représentatives de la catégorie, contenant les valeurs les plus fréquentes pour chaque caractéristique. Par exemple, pour le Tableau 1.1, ces prototypes se définiraient comme 0000 pour la catégorie A

[11] Dans cette structure, le fait que les exemplaires d'une même catégorie partagent un grand nombre de valeurs pour les caractéristiques explique l'utilisation du terme « ressemblance familiale ».

et 1111 pour B. Lorsque l'on utilise les symboles, ces prototypes sont des instanciations possibles, mais n'ayant jamais traitées par le système.

Lorsque les caractéristiques sont intégrales et difficilement identifiables (on parle alors de formes et catégories mal définies; *ill-defined* : Homa, 1978), le prototype peut alors être vu comme une simple moyenne arithmétique de tous les exemplaires d'une catégorie[12], calculé d'après une mesure de similarité globale. Le résultat expérimental classique soutenant l'existence de ce type de tendances centrales dans le système cognitif fut fourni par Posner et Keele[13] (1968), qui ont utilisé des stimuli perceptuels sans signification préalable. Ces auteurs ont montré qu'après avoir été exposés de façon répétée à un ensemble de versions distordues de nuage de points aléatoires prototypiques représentant chaque catégorie à l'étude, les participants humains démontraient la connaissance d'une abstraction générique similaire au prototype de départ (qui n'avait jamais été présenté). De plus, la mémoire pour cette abstraction générique était plus durable que celle des autres exemplaires vus à l'apprentissage[14] (Posner et Keele, 1970), et cet avantage était plus marqué pour les catégories de grande taille (Knapp et Anderson, 1984).

Ces résultats classiques ont été reproduits et étendus à de nombreuses occasions au cours des 40 dernières années (Homa, 1978; Homa, Rhoads, et Chambliss, 1979; Homa, Sterling, et Trepel, 1981; Marsolek, 1995; Smith, Redford et Haas, 2008; ainsi que plusieurs autres), utilisant les même stimuli originaux ou d'autres variantes catégorielles construites autour d'un prototype. D'autres résultats expérimentaux connexes basés sur des structures symboliques (Reed, 1972, 1978; Smith et Minda, 1998; entre autres), utilisant par exemple des visages schématiques ou des animaux artificiels comme stimuli, ont validé l'approche en montrant que l'exactitude des décisions catégorielles pour des exemplaires spécifiques était directement liée à leur similarité avec leur prototype prédéfini.

---

[12] Le prototype ainsi défini peut ne pas correspondre à une instanciation plausible.

[13] Même si, en fait, les stimuli qu'ils ont utilisés peuvent difficilement être traduits en une liste de symboles.

[14] Il est souvent rapporté dans la littérature que les expériences de Posner et de ses collègues constituent une preuve de l'adéquation d'une vision strictement prototypiste de la catégorisation. Cependant, les auteurs originaux ne prennent aucune position à ce sujet, se limitant à conclure à l'existence d'abstractions génériques sous forme de prototypes dans le système cognitif, et de l'importance d'en tenir compte dans les théories de catégorisation, sans nettement rejeter la possibilité de l'utilisation catégorielle d'une mémoire d'exemplaires.

La vision prototypiste a permis plusieurs avancées théoriques importantes. En premier lieu, elle a permis d'appliquer une définition opérationnelle de la similarité en termes de distance entre un stimulus et une représentation (Shepard, 1987), optique largement inspirée de la vision métrique multidimensionnelle des paysages cognitifs de Shepard (1958, 1962). En second lieu, elle a défini l'appartenance catégorielle comme un continuum, où un item peut être simultanément représentatif à différents degrés de plusieurs catégories (Murphy, 2002). Ainsi, l'inconstance de certaines décisions catégorielles (McCloskey et Glucksberg, 1978), ainsi que les différences temporelles dans leur application (Posner et Keele, 1968; Rips et al., 1973), peuvent être expliquées par la similarité quasi-équivalente entre un exemplaire donné et deux prototypes catégoriels. Finalement, en mettant l'accent principalement sur l'utilisation de matériel expérimental de type perceptuel[15], les tenants de l'approche prototypiste ont permis de nous éloigner du monde langagier, et d'ainsi explorer le monde de la formation implicite des catégories, reposant sur un algorithme mathématique pour les prises de décision catégorielles plutôt que sur le traitement logique d'un ensemble de règles verbales (Murphy, 2002).

#### 1.2.2.2 L'approche exemplariste

Malgré ses multiples avantages, l'approche prototypiste a éventuellement été jugée incomplète par certains, sur la base de résultats expérimentaux montrant une sensibilité humaine à la variabilité entre les exemplaires d'une catégorie (*e.g.*, Tversky et Kahnemann, 1973). Ces résultats, accompagnés d'une intuition selon laquelle la récupération consciente d'une catégorie ramène nécessairement à un exemplaire connu (Brooks, 1978), ont amené certains cognitivistes à proposer une nouvelle approche symbolique probabiliste basée sur une similarité comparative envers la totalité des membres d'une catégorie. Selon les tenants de cette approche dite « exemplariste » (Brooks, 1987; Medin et Schaffer, 1978; Nosofsky, 1986), on ne peut réduire une catégorie à une représentation abstraite, puisque sa formation et son utilisation sont des propriétés émergentes liées à la mémorisation et à l'utilisation de la totalité des exemplaires.

---

15 Ceci ne constitue pas une nouveauté en soi, compte tenu de l'utilisation de ce type de matériel par Shepard, Hovland et Jenkins (1961) pour l'étude des catégories booléennes.

Tout comme dans l'approche prototypiste, les décisions catégorielles sont définies par un calcul de similarité. Par exemple, dans le « Context Model » (Medin et Schaffer, 1978), l'exemplaire à classer est comparé à tous les exemplaires des catégories existantes dans le système représentationnel. Un coefficient de similarité total est calculé entre l'exemplaire et chacune des catégories. La règle de décision veut que l'exemplaire soit classé dans la catégorie pour laquelle ce coefficient est le plus élevé. Cette règle, la « règle des choix », tient compte des poids attentionnels rattachés à chaque dimension.

Les avantages de l'approche sont nombreux. En plus de répondre aux problèmes causés par les effets de typicité, la disjonctivité et l'inconstance décisionnelle, cette vision a produit des modèles permettant une remarquable adéquation quantitative aux données empiriques colligées lors de tâches de catégorisation. Le modèle le plus célèbre de l'approche demeure le Generalized Context Model (GCM : Nososfky, 1986, 1988; voir aussi l'implémentation connexioniste localiste, ALCOVE : Kruschke, 1992). Le GCM se veut une version générale du modèle de Medin et Schaffer (1978). Il prend en compte la similarité inter-exemplaires, ainsi que de multiples composantes cognitives comme la division du champ attentionnel décisionnel, le biais décisionnel préalable envers une catégorie donnée, ou la sensibilité à la fréquence de présentation des exemplaires. Le GCM permet également de modifier la métrique de distance utilisée, en fonction du type de dimensions (séparables ou intégrales) composant les stimuli.

Outre la performance quantitative spectaculaire du modèle[16], Nosofsky et son GCM ont permis d'enrichir la théorie générale du traitement d'objets et de la catégorisation. Entre autres, Nosofsky (1986) a démontré que l'on pouvait prédire les décisions catégorielles à l'aide des réponses obtenues à une tâche d'identification absolue de stimuli, suggérant un lien direct entre le traitement de l'objet au niveau individuel et la formation de groupement d'objets. Aussi, son insistance à analyser les performances au niveau individuel a permis de déceler les lacunes de la mise en commun de résultats de groupe, qui peuvent parfois cacher des différences de stratégie et de rapidité d'apprentissage. Finalement, son refus d'utiliser des définitions symboliques prédéterminées pour plutôt utiliser comme intrant du modèle les

[16] Plusieurs auteurs (dont Komatsu, 1992) estiment que la performance supérieure du GCM est en fat peu surprenante, par principe, puisqu'un modèle utilisant la totalité de l'information disponible devrait en principe mieux performer qu'un modèle n'utilisant qu'une partie de cette information.

coordonnées d'une analyse multidimensionnelle effectuée pour chaque participant (Nososky, 1986; Shin et Nosofsky, 1992) rappelle l'importance de ne pas considérer les paysages catégoriels comme uniques, ainsi que celle de tenir compte de la fréquence de présentation des exemplaires (Nosofsky, 1988) sur la composition de ces paysages.

De plus, alors que le succès des modèles prototypistes repose sur la séparabilité linéaire de l'espace multidimensionnel des stimuli, Medin et Schwanenflugel (1981; Wattenmaker, Dewey, Murphy, et Medin, 1986), utilisant des structures catégorielles symboliques, ont montré que les participants pouvaient apprendre des structures séparées linéairement ou non-linéairement, sans impact sur la performance. Les modèles exemplaristes pouvaient rendre compte de ce phénomène.

Cela dit, l'approche exemplariste demeure difficile à justifier, principalement parce qu'elle est très peu économique cognitivement (Goldstone et Kersten, 2003; Komatsu, 1992), vu le nombre de calculs requis, et surtout la quantité d'information à stocker en suivant le principe d'encodage obligatoire des exemplaires (Nosofsky et Palmeri, 1997)[17]. Aussi, certains défenseurs de l'approche prototypiste ont clairement montré que la grande majorité des résultats soutenant l'approche exemplariste ont été obtenus à l'aide d'une même structure catégorielle (la structure « 5-4 »; Medin et Schaffer, 1978) mal définie et peu plausible en termes du monde réel (Anderson, 1991; Minda et Smith, 2002; Smith et Minda, 2000). Finalement, il a été démontré que les effets supportant l'approche étaient généralement obtenus lorsque les expériences utilisent peu d'exemplaires présentés un grand nombre de fois aux participants (Murphy, 2002). Ceci gonflerait artificiellement la propension des participants à mémoriser chaque exemplaire pour réussir la tâche catégorielle, et rendrait en fait l'identification et la reconnaissance des exemplaires aussi facile que leur classification (Blair et Homa, 2003). Ainsi, un consensus croissant vise à affirmer que ces expériences montrent en fait la présence d'une certaine composante de mémoire épisodique d'exemplaires (ou « d'empreintes de stimuli »), sans toutefois convaincre du bien-fondé de l'utilisation de la totalité des représentations mnésiques pour catégoriser.

[17] Ce principe a d'ailleurs été mis en doute par une série d'expériences réalisées par Giguère, Lacroix et Larochelle (2007; Lacroix, Giguère et Larochelle, 2005), qui ont montré que les participants aux tâches de catégorisation n'étaient pas en mesure de mémoriser l'ensemble des associations inter-attributs dans un ensemble d'exemplaires, ce qui aurait en fait constitué une preuve inébranlable de la validité du principe d'encodage obligatoire.

Finalement, au cours des dernières années, plusieurs auteurs (Blair et Homa, 2001; Smith, Murray et Minda, 1997) ont ré-exploré le problème de la séparabilité. Ils ont conclu qu'il existe en fait pour les humains une contrainte de séparabilité linéaire influençant les performances de catégorisation perceptuelle. La clé pour démontrer cette limite cognitive repose sur l'utilisation de catégories mieux différenciées et plus réalistes, d'un plus grand nombre de catégories, et surtout, d'un nombre accru d'exemplaires. De façon plus générale, Thorpe, O'Regan et Pouget (1989; voir aussi Bourne, 1970) ont montré qu'au niveau perceptuel, les humains étaient incapables d'abstraire des règles de type « OU exclusif » (*XOR*). Pour ce faire, ils ont présenté sur de courtes périodes à leurs participants des grilles de carrés dont la présence suivait des règles simples, conjonctives ou disjonctives, que les participants devaient déterminer. Les participants ont réussi à identifier les règles simples et conjonctives, mais jamais les règles disjonctives.

### 1.2.3 Catégorie vs. objet : des systèmes distincts?

Alors que la « bataille des représentations » faisait rage au début des années 90, une théorie des systèmes multiples a été proposée pour rendre compte de processus distincts dans le système cognitif. Knowlton et Squire (1993; Knowlton, Mangels et Squire, 1996) ont montré, à l'aide de populations présentant des déficits neurologiques, l'existence de deux systèmes cognitifs séparés. Entre autres, ils ont demandé à des participants contrôle et des sujets amnésiques de catégoriser un ensemble de stimuli construit autour de prototypes (Posner et Keele; 1968, 1970) et d'effectuer une tâche de reconnaissance utilisant ces mêmes stimuli. Knowlton et Squire (1993) ont montré que la performance de reconnaissance était atteinte chez les amnésiques, mais pas la performance de catégorisation au test, alors que chez les sujets atteints de Parkinson, le résultat inverse fut observé (Knowlton, Mangers et Squire, 1996). Ainsi, ils ont conclu que les processus liés au niveau objet et niveau catégorie devaient reposer sur des systèmes de traitement et de représentation distincts, contrairement à l'hypothèse originale de Nosofsky (1986). L'un de ces systèmes effectuerait la catégorisation en développant des représentations prototypiques (ou basées sur les informations statistiques communes des exemplaires), et l'autre s'occuperait des tâches de reconnaissance de stimuli (tâche qui, rappelons-le, est intimement liée à celles d'identification absolue et de discrimination).

D'autres études empiriques conduites dans les années 90 ont également supporté une telle dissociation. Marsolek (1995) a proposé une étude basée sur la méthode de Posner et Keele, mais utilisant des stimuli prototypiques différents. Au rappel, il a présenté les exemplaires à classer soit à l'extrême gauche de l'écran (champ visuel lié à l'hémisphère droit) ou à l'extrême droite (hémisphère gauche). Il a pu montrer que les effets de prototype habituels (avantage pour les prototypes jamais vus, performance en fonction de la distance entre un nouvel item et son prototype) n'étaient présents que si l'on exposait les stimuli du côté droit de l'écran. Ces résultats montrent donc une dissociation au niveau de la catégorisation entre les deux hémisphères. Précédemment, utilisant l'identité visuelle de mots et une tâche d'amorçage perceptuel (*perceptual priming* : Marsolek, Kosslyn et Squire, 1992), il avait montré la présence d'un système faisant l'acquisition d'information perceptuelle spécifique, jugé plus performant lorsque lié au champ visuel gauche. Ainsi, Marsolek et ses collègues ont aussi montré, à l'aide d'un paradigme expérimental original, une dissociation au niveau implicite entre un présumé système basé sur les formes visuelles abstraites (niveau catégorie) et un autre basé sur l'information spécifique à chacun des exemplaires (niveau objet). Ces résultats ajoutent de la validité à l'hypothèse de Knowlton et de ses collègues.

Ainsi donc, si la catégorisation peut être effectuée sans aucune mémoire d'exemplaires, c'est qu'elle peut difficilement reposer sur la récupération de ceux-ci lors du test. Quelques indices empiriques préalables de cette conclusion avaient été présentés au cours des années 70, mais n'ont pas eu l'impact escompté. Hayes-Roth et Hayes-Roth (1977) ont mesuré de faibles corrélations entre les jugements de confiance dans des tâches de reconnaissance et de catégorisation utilisant les mêmes items. Reed (1978) a demandé à des participants d'effectuer simultanément (ou en alternance) des tâches d'induction catégorielle et d'identification utilisant le même ensemble de stimuli. Il a montré que les participants à cette double tâche apprenaient l'appartenance catégorielle beaucoup plus rapidement que l'étiquette d'identification. En fait, le processus d'identification ne pouvait débuter que lorsque l'apprentissage catégoriel était dans un stade avancé. En général, ces résultats empiriques mettent en doute le fait que les performances dans les deux types de tâches utilisent les mêmes traces mnésiques d'exemplaires et le même système de traitement,

puisque des mémoires d'exemplaires assez développées pour la catégorisation devraient également servir en reconnaissance et en identification (Smith et Minda, 2001)

Plusieurs exemplaristes, Nosofsky en tête, ont tenté de discréditer les résultats de Squire et collègues à l'aide de manipulations expérimentales et d'élaborations théoriques basées sur des modèles exemplaristes comme le GCM (Nosofsky et Zaki, 1998; voir aussi Palmeri et Flanery, 1999). Cependant, malgré toutes les défenses possibles basées sur des modèles, les résultats neurologiques soutenant la dissociation sont difficiles à ignorer. Reber, Stark et Squire (1998a; 1998b) ont différencié la localisation neurologique des processus cognitifs liés à la catégorisation et la reconnaissance à l'aide de tâches utilisant des distorsions de points. Durant une phase-test de catégorisation, ils ont montré une activité réduite dans l'aire occipitale postérieure pour les catégories apprises. Cette baisse d'activation n'était pas présente pour la tâche de reconnaissance; on pouvait plutôt y observer une hausse de l'activité dans le cortex occipital droit (les zones occipitales montrant les changements étaient clairement distinctes). Aussi, la découverte d'un patient montrant une amnésie antérograde extrêmement grave (patient E.P.) ont permis d'infirmer des hypothèses de travail basées sur une mémoire déclarative résiduelle (Hamann et Squire, 1997; voir aussi Reed, Hamann, Stefanacci, et Squire, 1997). Ce patient, lorsqu'exposé au même stimulus à 40 reprises consécutives, ne pouvait déterminer si ce patron avait été vu lors d'une tâche de reconnaissance. Cependant, ces capacités d'abstraction de prototypes étaient intactes. En comparaison, des participants contrôles réussissaient les deux tâches (Squire et Knowlton, 1995).

De nos jours, cette division entre des systèmes séparés traitant le niveau objet et le niveau catégorie est plutôt prise pour acquise par la plupart des théoriciens de la catégorisation (à l'exception des exemplaristes bien sûr). Plusieurs auteurs la prennent en compte et l'incorporent dans leurs modèles (voir ATRIUM : Erickson et Kruschke, 1998; COVIS : Ashby, Alfonso-Reese, Turken et Waldron, 1998; Mixture model : Smith et Minda, 1998 ; SPC : Ashby et Waldron, 1999; pour ne nommer que ceux-ci).

## 1.3 Extraction de caractéristiques

Alors que l'apport des théories et modèles symboliques à la compréhension de la catégorisation humaine est indéniable, deux problèmes majeurs subsistent toujours pour cette classe de modèles. Ces problèmes montrent l'incomplétude de l'approche symbolique, et la nécessité de compter sur la contribution de systèmes dits associationistes pour rendre compte des processus de catégorisation.

Le premier problème concerne le mutisme des théoriciens en ce qui a trait au processus de formation de catégories. En effet, la grande majorité des modèles proposés (à l'exception des modèles d'Ashby et collègues, reposant sur des procédures perceptuelles) n'expliquent les résultats observés que par l'utilisation des catégories déjà formées. Avec ces modèles, les mémoires d'exemplaires (Nosofsky, 1986) et de prototypes (Smith et Minda, 1998) sont encodées avant la première prédiction d'appartenance catégorielle. Ainsi, même si les modèles proposés constituent d'excellents outils permettant de déterminer le niveau d'utilisation des dimensions par les participants, ils ne nous proposent pas d'hypothèses à propos du mécanisme de développement des catégories.

Le second problème découle du premier, et est encore plus central. Selon Schyns et Rodet (1997), la quasi-totalité des théories de catégorisation laissent de côté le problème de la création et du développement des caractéristiques composant les stimuli. L'une des stratégies populaires est de pré-établir un « vocabulaire » de caractéristiques séparables (et théoriquement indépendantes), pour ensuite expliquer les processus cognitifs en termes d'opérations sur les éléments de ce vocabulaire (Goldstone, Schyns et Medin, 1997). Ainsi, les théoriciens et expérimentateurs, pour des raisons de contrôle expérimental, ont tendance à utiliser des stimuli simples, composés de caractéristiques saillantes, hautement distinctives, et verbalement identifiables (Schyns et al., 1998, parlent de stimuli qui « portent leurs caractéristiques sur leurs manches »). Ils peuvent même parfois indiquer aux participants la nature des caractéristiques qu'ils devront utiliser, lorsque cela n'entrave pas le but de l'expérience ou l'interprétation des résultats. Cet accord implicite rend plus facile la détermination du type de représentation mentale utilisé par les participants (Schyns et Rodet, 1997), qui est, tel que déjà mentionné, encore et toujours la principale source de désaccord dans le domaine.

Les modélisateurs suivent également cette tendance : pour les modèles symboliques, on considère les caractéristiques comme étant intrinsèques au système cognitif. Les listes de symboles constituent des définitions perceptuelles fixes, toujours de nature explicite, et donc prédéterminées par le modélisateur. Donc, par convention, un traitement arbitraire est implicitement appliqué aux intrants du système perceptuel, pour qu'un modèle à caractère symbolique puisse ensuite les traiter directement. Cet *a priori* peut sembler raisonnable lorsque les instructions expérimentales impliquent un accord clair sur l'identité des caractéristiques composant les stimuli. Cependant, dans les cas d'induction catégorielle sans instructions explicites quant à la nature des caractéristiques, il est plutôt difficile de soutenir cette méthode. En fait, dans cette situation empirique beaucoup plus commune, tout participant pourrait utiliser des caractéristiques plus simples ou plus complexes que celles ayant été théorisées. D'ailleurs, le but du participant dans une tâche de catégorisation n'est pas d'identifier les caractéristiques désirées ou de s'y conformer, mais bien de développer un ensemble de représentations perceptuelles diagnostiques permettant d'accomplir la tâche (Goldstone et al., 1997).

Il existe d'ailleurs plusieurs domaines où les caractéristiques définissant les objets s'avèrent complexes, nettement intégrales, inconnues préalablement à l'apprentissage. Schyns et Murphy (1994) relèvent en outre quelques tâches où c'est le cas, tels l'identification de bactéries au microscope, ou encore la lecture de radiographies par les médecins. Apprendre à interpréter des radiographies implique non seulement de pouvoir apprendre l'identité des caractéristiques, mais bien de les extraire de l'ensemble de stimuli, et de les associer à un diagnostic connu. Ainsi, la communication de l'expertise diagnostique entre le médecin expert et l'étudiant en médecine est ardue, à cause du caractère complexe et intégral des composantes perceptuelles (Brooks, Norman , et Allen, 1991). Le novice doit plutôt développer son propre système d'interprétation par apprentissage inductif, et ce système sera différent pour chaque individu.

Gibson (1969) argumentait d'ailleurs en ce sens il y a longtemps, en affirmant que l'interprétation perceptuelle de l'environnement est « personnelle », et dépend donc de l'historique et de l'entraînement de l'observateur. Conséquemment, il est raisonnable de croire qu'en général, les représentations des dimensions perceptuelles pourraient être

apprises; elles se développeraient à mesure que le système cognitif abstrait l'information statistique. Goldstone, Schyns et Medin (1997) opinent également dans ce sens : selon eux, les chercheurs du domaine perceptivo-cognitif devraient toujours tenir compte des processus d'apprentissage perceptuels, puisqu'ils expliquent en grande partie la variance de performance dans les tâches expérimentales. Schyns et Murphy (1994) croient qu'il est simpliste de croire que les caractéristiques qui forment la base du processus de catégorisation sont universelles et clairement « données » par l'environnement.

Malgré le grand intérêt du sujet, peu d'auteurs se sont attaqués à ce difficile problème du développement des caractéristiques. Il existe toutefois quelques preuves empiriques montrant que le système perceptuel est flexible, et que les caractéristiques sont ajustées selon l'expérience perceptuelle et l'historique catégoriel de chaque individu (Schyns et Rodet, 1997). Également, certaines expériences ont montré le caractère complexe des caractéristiques visuelles, caractère permettant difficilement la transformation en termes verbalisables, par exemple.

Schyns et Murphy (1991; 1994) ont présenté à leurs participants des ensembles de « roches martiennes » tridimensionnelles qu'ils devaient associer à des étiquettes catégorielles. Ces stimuli étaient hautement riches et complexes, et ne contenaient aucune dimension verbalisable ou particulièrement saillante. La richesse fut démontrée par une tâche pré-expérimentale, où les participants montraient un faible niveau d'accord sur l'identité des composantes pertinentes. Les catégories étaient définies par un petit groupe cohérent de « blobs » contigus sur chacun des stimuli. Suivant la tâche, les participants devaient utiliser un programme pour décomposer et indiquer, à l'aide d'un curseur à l'écran, quelles caractéristiques diagnostiques ils avaient utilisé pour effectuer la tâche. Les résultats montrèrent que malgré un taux de réussite élevé chez tous les participants, ces derniers n'utilisaient pas les mêmes parties d'objet pour catégoriser. Aussi, Schyns et Murphy ont trouvé que l'apprentissage catégoriel en était venu à modifier le vocabulaire de caractéristiques perçues par les participants, ce vocabulaire étant donc nettement flexible et adaptif.

Schyns et Rodet (1997), dans une série d'expériences de catégorisation de « cellules martiennes », ont également fourni plusieurs effets intéressants. Ces stimuli différaient des

roches martiennes en ce qu'ils étaient en deux dimensions, et contenaient des caractéristiques saillantes de grande taille. Elles étaient toujours non-verbalisables. Schyns et Rodet ont montré, dans une tâche préalable à leur première expérience, que la simple pré-exposition aux stimuli était suffisante pour que les participants identifient des composantes de cellules leur semblant significatives, ceci constituant une preuve de la présence d'un mécanisme autonome de création de caractéristiques. À l'aide d'expériences de catégorisation utilisant ces cellules, ils ont aussi montré que l'ordre d'apprentissage des catégories pouvait influencer le vocabulaire de caractéristiques orthogonales développé par les participants, nous ramenant une fois de plus à l'importance de tenir compte de l'historique perceptuel individuel. Finalement, leurs résultats ont permis de montrer que le système perceptuel n'extrait pas nécessairement les primitives perceptuelles les plus simples, mais plutôt les primitives suffisamment simples pour permettre la différentiation.

Ces tâches sont extrêmement ardues à modéliser, puisque les systèmes symboliques existants manquent de flexibilité. Ils sont incapables de créer de façon autonome les symboles qu'ils traitent. Les modèles courants sont sur-contraints par l'utilisation de primitives trop abstraites, et inflexibles par leur manque d'interactions avec l'environnement perceptuel (Goldstone et al., 1997; Harnad, 1990). Par conséquent, si une catégorisation précise repose sur une caractéristique inexistante, alors la tâche ne peut être réalisée (Schyns et al., 1998). Ainsi, une théorie complète de la catégorisation ne devrait pas se limiter à tenter de déterminer la méthode de recombinaison des caractéristiques composant les stimuli, mais devrait aussi expliquer l'origine et le développement de ces caractéristiques, puisqu'une partie significative de l'apprentissage catégoriel repose sur leur détermination (Schyns et Murphy, 1994; Schyns et Rodet, 1997; Schyns et al., 1998).

Harnad (1990) propose donc qu'une architecture cognitive complète devrait être en mesure de créer et utiliser son propre « code » sans intervention ou prétraitement. Pour ce faire, il propose une structure en trois niveaux, les deux premiers étant non-symboliques et basés sur l'induction et le traitement ascendant. Selon lui, un modèle émanant de cette structure devrait en premier lieu utiliser un réseau de neurones artificiels pour extraire un ensemble de représentations « iconiques », utiles pour la différentiation au niveau objet. Edelman et Traitor (1997) sont d'avis que ce traitement des objets ne peut se faire de façon

brute, puisque le nombre d'exemplaires nécessaires pour l'apprentissage augmente de façon exponentielle en fonction du nombre de dimensions[18]. Ainsi, une réduction dimensionnelle lors du traitement, une compression de l'entrée visuelle, tombe sous le sens, et devient alors la véritable base des processus de généralisation dans l'espace multidimensionnel (Shepard, 1987). Ces réductions dimensionnelles n'ont pas besoin d'être explicites, mais doivent strictement être définies par leur rôle fonctionnel dans l'apprentissage perceptuel et la catégorisation (Schyns et al., 1998).

Ensuite, Harnad (1990) affirme que les caractéristiques iconiques extraites devraient être transformées en représentation catégorielles, contenant les caractéristiques invariantes des projections sensorielles permettant de distinguer les membres des non-membres d'une catégorie. Finalement, pour compléter le système cognitif, ces représentations devraient selon Harnad être transformées par un protocole quelconque, en patrons structurés que sont les symboles, lorsque la situation le permet. La manipulation de symboles pourrait alors être réalisée pour les décisions basées sur une mémoire explicite.

Schyns, Goldstone et Thibaut (1998; Goldstone, 1998) ont parfaitement supporté cette vision, en militant pour des modèles cognitifs autonomes pouvant créer un ensemble de « blocs de construction perceptuels », qui pourront ensuite être réutilisées pour diverses tâches perceptuelles effectuées par le système. Leur théorie soutient que le répertoire de caractéristiques n'est pas fixe, mais repose sur les demandes de la tâche, les nécessités catégorielles et les contingences environnementales. Goldstone, Schyns et Medin (1997) croient qu'un tel système de développement de caractéristiques devrait être adaptif, lui permettant de modifier le type de caractéristiques extraites selon l'environnement, et de varier les instanciations des caractéristiques pouvant ainsi se mouler à différents domaines de catégorisation perceptuelle. Ainsi, ils sont d'avis que la flexibilité est cruciale dans un tel système (retour au principe de plasticité : Grossberg, 1987). Selon eux, les systèmes perceptuels devraient posséder des mécanismes généraux adaptifs leur permettant de développer les nouveaux outils nécessaires à la réussite de la tâche.

---

[18] Cette position théorique concorde avec celle, précedemment mentionnée, de Barlow (1961), pour qui le postulat de réduction dimensionnelle des entrées perceptuelles est central à la réussite du système perceptive-cognitif.

Plusieurs avantages pourraient découler de la proposition de tels modèles. Pour débuter, la flexibilité du système par apprentissage de nouvelles caractéristiques devrait réduire la nécessité théorique de devoir compter sur des règles complexes ou booléennes (Goldstone et al., 1997). Un tel système serait aussi plus parcimonieux que l'approche symbolique présente, puisqu'il permettrait de ne mémoriser que les composantes servant la perception et la catégorisation. Finalement, alors que dans un système à dimensions fixes et prédéterminées, la plupart des caractéristiques seraient peu utilisés, les modèles capables d'extraire les composantes selon les besoins pourraient en fait offrir une adéquation quasi-optimale, une adaptation maximale à l'environnement et aux domaines perceptuels traités (Schyns et al., 1998).

## 1.4 Conclusion

Le présent chapitre a exploré les problématiques de base en apprentissage et catégorisation perceptuels. Une conclusion raisonnable, découlant des nombreuses théories présentées, est qu'à la base, un modèle perceptivo-cognitif devrait viser à extraire les informations statistiques de l'environnement, dans le but d'instancier certaines composantes complexes et de réduire la dimensionnalité de l'entrée perceptuelle. Cette extraction devrait graduellement mener à une meilleure discrimination parmi les différents objets et les différentes catégories.

En tenant compte des résultats de Knowlton, Squire et collègues, il semble également qu'un tel modèle de catégorisation devrait être en mesure de créer les représentations d'objet et de catégories dans des espaces multidimensionnels cognitifs séparés (double dissociation et théorie des systèmes multiples). Aussi, des principes connexes devraient être utilisés dans le but de mener à des processus de traitement autonomes, et d'autres processus d'apprentissage supervisé de contingences entre les représentations d'objets ou de catégories apprises par le système, et de réponses prédéterminées.

Le prochain chapitre détaillera les nombreuses hypothèses de travail découlant de cette recension des écrits, et tentera de déterminer l'adéquation de différentes classes de modèles de réseaux de neurones pour chacune des hypothèses, de façon globale.

# CHAPITRE II

## LE CHOIX D'UNE ARCHITECTURE ADAPTÉE AUX OBJECTIFS DU PROJET

Suite à la recension des théories et modèles présentée dans le chapitre précédent, il est possible d'énumérer des caractéristiques générales désirables pour un modèle perceptivo-cognitif. Le présent chapitre mettra l'accent sur ces caractéristiques, ainsi que sur les tâches qu'un tel modèle devrait être en mesure d'effectuer. Après avoir détaillé les caractéristiques et tâches, et brièvement argumenté quant au choix d'utiliser les réseaux de neurones artificiels, plusieurs classes de modèles de ce type seront décrites et évaluées.

Le modèle de base présenté dans cette thèse en sera un d'apprentissage de type perceptif. Le modèle se déclinera en deux versions, soit autonome (Chapitre 3 et 4) ou à rétroaction externe (Chapitre 5). En accord avec les principes de Gibson (1969), lorsque le modèle effectuera de l'apprentissage autonome, son but principal sera la différentiation entre les objets composant l'environnement. De façon optimale, une « expertise perceptuelle » se développera progressivement : ainsi, le modèle qui vise la différentiation pourra graduellement reproduire les stimuli de façon de plus en plus précise. Il est à noter que la différentiation parfaite est un but théorique; l'inclusion de ce critère technique dans le modèle n'implique nullement que les humains visent à pouvoir reproduire sans failles les stimuli qui frappent leurs rétines. En fait, tel que précédemment mentionné, le but de l'apprentissage perceptuel est plutôt la mise en place de traces mnésiques suffisamment distinctes.

En accord avec la théorie de Shepard (1958, 1987), le modèle devra aussi être basé sur une approche multidimensionnelle de la cognition. Ainsi, tel que généralement accepté par les théoriciens en psychologie cognitive (voir Ashby, 1992; Nosofsky, 1992; Perrin, 1992;

entre autres), les représentations de stimuli devront être encodés dans le système comme des points dans un espace multidimensionnel, basé sur une métrique de distance prédéfinie[19].

Finalement, suivant les propositions de Goldstone (1998), Goldstone, Schyns et Thibaut (1998), Hall (1991) et Harnad (1990), l'apprentissage devra se faire de façon ascendante, c'est-à-dire que le modèle devra obtenir son information de l'environnement perceptuel. Ainsi, il devra partir du niveau « pixel » pour créer ses représentations internes de stimuli et de catégories. Lorsqu'il fonctionnera de façon autonome, aucune autre information ne lui sera disponible.

Ces mêmes auteurs ont tous proposé que les processus décrits devraient être réalisés à l'aide de réseaux de neurones artificiels. Ces modèles constituent d'excellents outils pour modéliser l'apprentissage ascendant. Aussi, ils ont généralement beaucoup de succès dans la modélisation de processus implicites de bas niveau (liés à l'apprentissage perceptuel). Par définition, les réseaux de neurones visent à trouver les séparations optimales (droites, plans, hyperplans) dans l'espace permettant de résoudre un problème donné (Ashby & Gott, 1988). Contrairement aux approches symboliques, les modèles neuronaux possèdent une propriété intrinsèque de généralisation liée aux approches par similarité. Ils sont donc adéquats pour répondre à des problématiques de différentiation graduelle au niveau perceptuel. Comme il sera montré, ces modèles sont aussi adaptés à plusieurs autres tâches perceptivo-cognitives autonomes et supervisées.

Au cours du présent chapitre, une exploration de différentes classes de réseaux de neurones sera effectuée. Chaque classe sera évaluée en fonction de sa capacité à réaliser certaines tâches qui seront définies dans ce qui suit.

## 2.1 Tâches désirées

### 2.1.1 Réduction dimensionnelle

Tel que mentionné dans le Chapitre 1, de nombreux auteurs (*e.g.*, Barlow, 1961; Schyns et Rodet, 1997; pour ne nommer que ceux-ci) affirment que les stimulations perceptuelles,

[19] En général, on pense à la distance euclidienne pour les dimensions intégrales, ou la distance « city-block pour les dimensions séparables (Nosofsky, 1986).

avant de devenir des représentations mentales, doivent passer par un processus de réduction dimensionnelle. Cette compression de l'information rendrait les stimuli plus faciles à encoder, et permettrait donc une économie cognitive au niveau de la taille des représentations. Des représentations plus comprimées permettront ensuite une récupération et un traitement décisionnel plus rapides. Un modèle perceptivo-cognitif devrait donc être en mesure de produire une économie représentationnelle, en passant d'un mode « pixels » à un code plus restreint développé de façon autonome. Le réseau devra donc créer des représentations économiques, et associer chaque vecteur d'entrée (stimulation perceptuelle) à sa représentation comprimée dans l'espace « psychologique ».

### 2.1.2 Extraction de caractéristiques

La grande majorité des psychologues cognitifs s'entendent pour affirmer que le traitement perceptivo-cognitif repose sur un vocabulaire de composantes, de caractéristiques (voir entre autres : Garner, 1974; Goldstone et al., 1997; Komatsu, 1992; Medin et Schaffer, 1978; Smith et Medin, 1981; Murphy, 2002). Il a été montré au Chapitre 1 que la nécessité de prédéfinir ces composantes constituait une grave lacune des modèles symboliques. Le réseau proposé devra donc être en mesure de créer son propre ensemble de caractéristiques iconiques, qui ne seront pas nécessairement interprétables (Harnad, 1990). Nous savons que dans certaines conditions (lorsque le vocabulaire contient des composantes à saillance très élevée, par exemple), les humains peuvent rapidement récupérer les caractéristiques utilisées pour définir les objets (et même dessiner les contours de ces caractéristiques : Schyns et Murphy, 1994; Schyns et Rodet, 1997). Il serait donc nécessaire pour ce modèle de pouvoir récupérer les caractéristiques sans devoir recourir à un quelconque processus d'analyse complexe.

### 2.1.3 Développement autonome d'une mémoire d'exemplaires

Suite aux travaux de Knowlton, Squire et de leurs collègues sur la dissociation entre le niveau objet et le niveau catégorie (Knowlton et Squire, 1993; Knowlton et al., 1996; Reber et al., 1998a, 1998b; Squire et Knowlton, 1995), on doit présenter un modèle pouvant créer une mémoire d'exemplaires, c'est-à-dire un espace où chaque stimulation perceptuelle est

liée à un point distinct dans l'espace. Tel que sera discuté, ceci sera la contrainte la plus facile à satisfaire, puisqu'une classe de modèles, les mémoires autoassociatives récurrentes, sont des mémoires d'exemplaires par défaut.

### 2.1.4 Catégorisation autonome

Le réseau, tout comme les humains, devra être capable de créer et mémoriser des représentations uniques pour des ensembles de stimuli perceptivement distincts. Dans ce cas, en comparaison avec la mémoire d'exemplaires, un point dans l'espace devra représenter plusieurs stimulations perceptuelles distinctes, mais qui montrent une similarité au niveau global. Ce traitement devra se faire sans aucune rétroaction externe. Par parcimonie, la procédure de développement des catégories et celle de développement de mémoire d'exemplaires devraient suivre les mêmes postulats de base, et ne différer que d'une manière quantitative.

### 2.1.5 Catégorisation/Identification à l'aide d'une rétroaction externe

Tous comme les humains dans des tâches de laboratoire, on voudra qu'une seconde version du modèle soit également capable de profiter de rétroaction externe en ce qui a trait à l'étiquette associée à une catégorie. On pourra ajouter au modèle autonome un second module permettant l'association des représentations avec des étiquettes identificatives ou catégorielles. Une fois de plus, par parcimonie, il serait intéressant de pouvoir produire deux versions du même modèle (autonome et avec rétroaction) basées sur les mêmes postulats de base. C'est ce qui sera donc tenté.

## 2.2 Introduction aux principales classes de modèles neuronaux

Les réseaux de neurones artificiels (RNA : Anderson, 1995; Freeman, 1994; Haykin, 1994; Hertz, Krogh et Palmer, 1991; Negnevitsky, 2002) constituent une métaphore basée sur le fonctionnement du cerveau, inspirée entre autres par les principes évoqués par Donald Hebb (1949; dans Haykin, 1994). À la base, ces modèles sont composés d'unités représentant des neurones connectés les uns aux autres par des liens auxquels des poids numériques sont

associés. Dans la plupart des cas, chaque unité ne contient qu'une partie de l'information à traiter, ce qui rend l'information distribuée et globale dans le modèle (Negnevitsky, 2002).

Parmi les architectures de RNA les plus populaires, on retrouve les réseaux de type « feedforward » (à simple couche ou « multicouches »), où l'information est propagée de l'entrée du réseau vers la sortie, et les réseaux autoassociatifs récurrents (ou à attracteurs), où l'information est mise à jour par le réseau jusqu'à l'atteinte d'un niveau de stabilité adéquat (Negnevitsky, 2002). Ces deux types de réseau sont illustrés à la Figure 2.1. Dans ces deux cas, un principe est commun : chaque unité d'entrée reçoit une activation (discrète ou continue) qui provient de l'extérieur du réseau. Ensuite, l'activation est transmise à travers les connexions vers les autres unités. L'activation des unités subséquentes est une combinaison (linéaire ou non) de l'activation des unités qui la précède et des poids des connexions les liant.

Il existe plusieurs façons de classifier les différents réseaux de neurones. Une façon fréquente de le faire est diviser les réseaux selon la présence ou l'absence de supervision à l'apprentissage. La prochaine section présentera les architectures de base de réseaux de neurones, dans le but d'évaluer leur caractère adéquat face à chacune des tâches énumérées. Les architectures hybrides, complexes, ou topologiques ne seront pas considérées par principe de parcimonie.

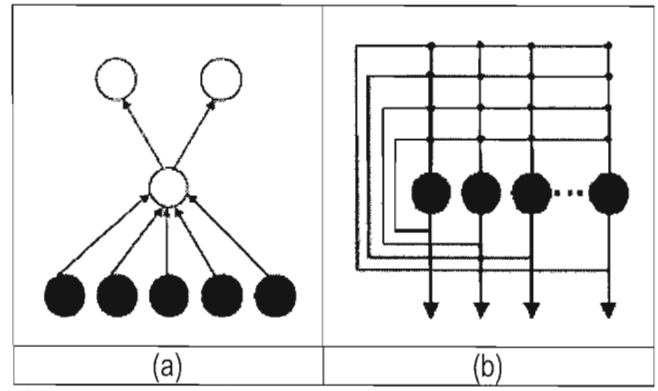


**Figure 2.1** Architectures connexionnistes les plus populaires. Les unités noircies sont les unités d'entrées. (a) Architecture de type *feedforward*. Toutes les connexions pointent vers l'avant et il n'y a pas de connexions entre les unités appartenant à une même couche. (b) Architecture récurrente. Chaque unité est connectée à toutes les autres (et dans certains cas à elle-même) et la sortie du réseau au temps $t$ constitue son entrée au temps $t + 1$.

### 2.2.1 Réseaux non-supervisés ou autonomes

#### 2.2.1.1 Règles d'apprentissage hebbiennes

*Mémoires autoassociatives récurrentes*

Les mémoires autoassociatives récurrentes (RAM) sont étroitement liées au concept de traitement autonome. Ces modèles sont basés sur la production d'une matrice de projection permettant aux entrées de converger dans un bassin d'attracteurs. Une caractéristique intéressante des RAM est l'utilisation d'une boucle de rétroaction, qui permet entre autres la reconnaissance de patrons d'entrée (*pattern recognition*), la généralisation à de nouveaux stimuli, la filtration du bruit, et la reconstruction de patrons incomplets (*pattern completion*) (Negnevitsky, 2002). La rétroaction interne permet à un réseau donné de passer progressivement d'un patron initial à un état invariable (en l'occurrence, un attracteur). Parce que l'information est distribuée entre les unités, les attracteurs du réseau ont la possibilité d'être globaux; si le réseau est entraîné correctement, ces attracteurs devraient correspondre aux stimuli appris (Hopfield, 1982).

Barlow (1989) considère que ce type de réseau est idéal pour l'apprentissage non-supervisé, parce qu'il utilise avantageusement la redondance contenue dans les signaux provenant du monde externe. L'une des façons de tenir compte de cette redondance est de compter sur une mesure de covariance entre l'activation des différentes unités. C'est exactement ce que propose l'apprentissage hebbien qui est à la base des RAM (Kohonen, 1972; Anderson, 1972)

Anderson, Silverstein, Ritz et Jones (1977) ont proposé le « Brain-State-in-a-Box » ou BSB, un modèle RAM dont les principes sont intéressants, mais qui souffre de plusieurs lacunes. Le principal problème est la nécessité d'utiliser des patrons d'entrée normalisés et orthogonaux entre eux[20]. Lorsque l'on utilise des patrons corrélés, le réseau est dominé par le premier vecteur propre et perd toute sélectivité. De plus, puisque que la règle d'apprentissage est strictement additive, avec la répétition des essais d'apprentissage, les poids de connexion croîtront à l'infini, provoquant « l'explosion du réseau ». Pour éviter ceci, Hopfield (1982) a

[20] Ces caractéristiques sont peu plausibles au point de vue de la validité écologique, puisque pour les humains, les entrées perceptuelles sont nettement corrélées au niveau des pixels.

proposé un réseau dont la matrice de poids de connexion représente la corrélation entre les entrées et elles-mêmes, suivant :

$$\mathbf{W} = \frac{1}{N}(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{I}) \qquad 2.1$$

où **X** représente la matrice contenant les patrons bipolaires d'entrée, *N* le nombre de patrons, **I** la matrice identité et **W** représente la matrice de poids de connexions. La matrice identité est utilisée afin de retirer les auto-connexions; caractéristique qui diminue les performances du réseau (Hopfield, 1982). La règle de transmission du réseau implique une fonction de type *signum,* définie comme :

$$signum(z) = \begin{cases} 1 \text{ si } z>0 \\ 0 \text{ si } z=0 \\ -1 \text{ si } z<0 \end{cases} \qquad 2.2$$

où *z* représente un nombre réel. Cette fonction oblige le réseau à ne créer des attracteurs qu'aux vertex de l'espace stimuli (qui peut être vu comme un hypercube). Ce type de réseau récurrent permet de traiter les stimuli corrélés à la condition que l'apprentissage soit unique, c'est-à-dire que chaque patron n'est vu qu'une seule fois. Par conséquent l'apprentissage est de type « hors-ligne », une solution peu plausible psychologiquement.

Quelques solutions ont été proposées pour remédier à la croissance des poids de connexions à l'infini. Une première solution fut de normaliser la matrice de poids de connexions selon un critère de projection optimale. Les mémoires associatives linéaires optimales (OLAM) permettent d'obtenir un spectre de valeurs propres égales; ainsi, aucun attracteur ne domine les autres. Initialement, la pseudoinverse fut utilisée (Personnaz, Guyon, et Dreyfus, 1985) mais elle nécessite l'utilisation d'une opération non-locale; conséquemment, l'apprentissage n'est plus de type hebbien (Brown, Kairiss et Keenan, 1990). Plusieurs auteurs (*e.g.*, Bégin & Proulx, 1996; Diederich et Opper (1987), Storkey et Valebregue (1999)) ont ainsi proposé des algorithmes permettant une convergence vers la solution de la pseudoinverse (ou à un facteur près) de façon itérative et locale. Dans tous les cas, la convergence ne peut s'obtenir sans facteur de correction (ou apprentissage anti-hebbien : Haykin, 1994) (Bégin et Proulx, 1996; Christos, 1996; Hopfield, Feinstein et Palmer, 1983). Malheureusement, certains de ces réseaux montrent plusieurs lacunes telles la normalisation nécessaire des stimuli, l'apprentissage hors-ligne ou la fixation de paramètres par essais et erreurs. Aussi, les stimuli doivent toujours être définis de façon bipolaire (ou

binaire) sans possibilité de développer des attracteurs ailleurs qu'aux coins d'un hypercube. Finalement, les espaces de solution de l'apprentissage corrélé contiennent beaucoup d'états (ou attracteurs) nuisibles ne figurant pas dans la banque initiale de stimuli. Ainsi, la taille des bassins d'attraction est moindre, ce qui réduit la performance au rappel.

Chartier et Proulx (2005) ont proposé un modèle permettant l'apprentissage corrélé, à procédures locales, tout en évitant l'explosion des poids de connexion. Pour ce faire, ils ont utilisé le concept d'apprentissage hebbien/anti-hebbien propre au modèle EIDOS (Bégin et Proulx, 1996), mais ont modifié la règle d'apprentissage et la règle de transmission. Le modèle résultant, NDRAM, utilise une nouvelle règle d'apprentissage en ligne prenant en compte la sortie et l'état initial du réseau :

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(\mathbf{x}(0)\mathbf{x}(0)^T - \mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}) \tag{2.3}$$

où $\mathbf{x}(0)$ représente le vecteur d'entrée initial, $\mathbf{x}(t)$ représente l'état final du vecteur d'entrée $\mathbf{x}$ après $t$ itérations dans le réseau, $\mathbf{W}$ représente la matrice de poids de connexions, et $k$ représente l'essai d'apprentissage. Le réseau utilise aussi une règle de transmission basée sur une correction de l'activation suivant une décroissance d'ordre trois :

$$\forall i,\ldots,N, \mathbf{x}_i(t+1) = \begin{cases} 1, \ Si\ \mathbf{W}\mathbf{x}_i(t) > 1 \\ -1, \ Si\ \mathbf{W}\mathbf{x}_i(t) < -1 \\ (\delta+1)\mathbf{W}\mathbf{x}_i(t) - \delta(\mathbf{W}\mathbf{x})_i^3(t), \ Sinon \end{cases} \tag{2.4}$$

où $N$ représente la dimensionnalité de l'entrée. La combinaison de ces deux règles permet à NDRAM de traiter des entrées non-normalisées et corrélées entre elles (représentées par des vecteurs à valeurs continues), et ce, en proposant une performance de rappel normal et bruité supérieure et une réduction substantielle du nombre d'attracteurs nuisibles.

Les RAM en général, bien que présentées comme des modèles de catégorisation, ne sont pas des modèles de *formation* de catégories. En général, lors de l'apprentissage, on fournit au réseau les prototypes de chacune des catégories; il n'a donc pas à les développer. Lors du rappel, on pourra bel et bien présenter de nouveaux stimuli, jamais traités par le réseau, et ces vecteurs pourront se stabiliser sur des attracteurs basés sur les prototypes (comportement de généralisation). On ne pourrait cependant tenter, lors de l'apprentissage, de développer des prototypes de catégories composées de vecteurs distincts à l'apprentissage (comme l'humain

le fait lorsqu'il crée des représentations catégorielles). Dans ce cas, la RAM créera un attracteur (un état) pour chaque vecteur distinct. Avec les essais d'apprentissage, l'espace-stimuli pourrait devenir identique à l'espace-réseau. En ce sens, les RAM constituent donc d'excellents modèles de création de mémoires d'exemplaires potentiellement reconstructibles de façon parfaite ou imparfaite.

Ces réseaux ne pourront pas non plus effectuer de réduction dimensionnelle, puisqu'ils ne contiennent qu'une couche d'unités; ainsi, le nombre d'unités de représentation est le même que celui représentant les entrées. Aussi, comme il n'y a qu'une couche, ceci rend impossible l'association directe entre une entrée perceptuelle et sa propre compression dans le réseau. Pour ce qui est de l'extraction de caractéristiques, on pourrait dire que les RAM en font. Cependant, pour récupérer ces caractéristiques, on doit subséquemment effectuer une analyse en vecteurs et valeurs propres, ce qui va à l'encontre des principes de simplicité énoncés en début de chapitre, et n'est possible que lors d'un apprentissage linéaire (ex. EIDOS: Bégin et Proulx, 1996). De plus, cette récupération est en fait externe, car les valeurs propres n'existent pas sous une forme implicite.

*Réseaux à base d'analyses en composantes principales*

Un second type de modèle utilisant l'apprentissage hebbien est celui des réseaux à base d'analyse en composantes principales ou PCA (Diamantaras et Kung, 1996; Foldiak, 1989; Oja, 1982; Sanger, 1989; Rubner et Schulten, 1990). Pour modéliser la stratégie de base voulue, en l'occurrence l'extraction de caractéristiques perceptuelles, une solution statistique évidente serait d'utiliser ce type de réseau. Les réseaux à base de PCA (Diamantaras et Kung, 1996) peuvent résoudre le problème de la variabilité des intrants en créant (ou « extrayant ») des caractéristiques statistiques de bas niveau représentant, jusqu'à un certain degré, l'information intrinsèque contenue dans les stimuli. Tout comme les RAM, les réseaux à base de PCA répondent au problème de la croissance infinie des poids de connexions grâce à un facteur de correction. De plus, afin d'extraire des composantes non redondantes, l'architecture est généralement contrainte par l'utilisation de connexions latérales hiérarchiques entre les unités de la couche de sortie (*e.g.*, Rubner et Schulten, 1990).

De façon générale, les réseaux de neurones à base de PCA, qui n'utilisent qu'une seule matrice de poids, sont basés sur une fonction de transmission linéaire définie comme suit:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{2.5}$$

où $\mathbf{x}$ représente le vecteur d'entrée original, $\mathbf{W}$ est la matrice de poids stabilisée, et $\mathbf{y}$ est le vecteur de sortie. Lorsque la matrice de poids mène à la préservation maximale des caractéristiques pertinentes, l'opération inverse:

$$\hat{\mathbf{x}} = \mathbf{W}^{\mathrm{T}}\mathbf{y} \tag{2.6}$$

devrait produire un vecteur résultant ressemblant maximalement au patron d'entrée original (principe de compression/reconstruction). Ainsi, le réseau doit minimiser la norme du vecteur représentant la différence entre le vecteur d'entrée et sa représentation comprimée correspondante. Ceci signifie trouver une matrice $\mathbf{W}$ qui minimise la fonction d'erreur:

$$E = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \hat{\mathbf{x}}_i \right\|^2 = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}_i \right\|^2 \tag{2.7}$$

où $E$ représente la fonction d'erreur à minimiser, et $N$ représente la dimensionnalité du vecteur d'entrée. Plusieurs solutions existent pour cette équation (pour une revue complète : Diamantaras et Kung, 1996), qui mène généralement vers une version dynamique du processus d'orthogonalisation de Gram-Schmidt.

La généralisation de ces équations au cas non-linéaire (nPCA : Karhunen, Pajunen et Oja, 1998) s'avère plutôt directe; dans ce cas précis, une fonction de transmission non-linéaire est utilisée :

$$\mathbf{y} = f(\mathbf{W}\mathbf{x}) \tag{2.8}$$

et la fonction de minimisation correspondante devient donc :

$$E = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{W}^{\mathrm{T}} f(\mathbf{W}\mathbf{x}_i) \right\|^2 \tag{2.9}$$

Différents types d'optimisation peuvent être réalisés selon le type de fonction non-linéaire utilisée. Si la fonction utilise un polynôme de degré 3 ou supérieur, le réseau est alors directement relié à l'analyse en composantes indépendantes (ICA) et aux autres algorithmes utilisés pour la séparation de signaux mixtes (Karhunen, Pajunen et Oja, 1998; Hyvarinen et Oja, 2000).

Alors que les réseaux de neurones à base de PCA sont fréquemment utilisés pour des tâches de compression d'images fixes (Sanger, 1989) et animées (Diamantaras et Kung, 1996), ces modèles n'ont pas su s'imposer dans le monde de la modélisation cognitive, même si certains possèdent de multiples propriétés intéressantes telles que des règles locales psychologiquement plausibles, et peuvent extraire de multiples composantes simultanément de façon parallèle. À la base, deux buts cruciaux de ces réseaux sont l'extraction de caractéristiques et la réduction dimensionnelle, tâches pour lesquelles une architecture est ici recherchée. L'utilisation de ce type de réseau ne nécessiterait pas d'analyse supplémentaire avant de récupérer les caractéristiques dans le réseau, puisque c'est exactement l'opération que le réseau accomplit.

Cela dit, la catégorisation (et la différentiation) autonome est généralement liée aux comportements de type « attracteur ». Ce type de comportement catégoriel ne peut être observé dans les réseaux d'extraction de composantes, puisque cette classe est unidirectionnelle (*feedforward*) par définition. En d'autres mots, les poids de connexions optimisent la projection des stimuli entre l'espace-réseau et l'espace-stimuli, mais aucun mécanisme ne permet de circuler dans l'espace-réseau vers un attracteur. Par conséquent, ces réseaux ne peuvent tolérer le bruit et l'incomplétude lors du rappel. Aussi, les postulats de base menant vers des composantes orthogonales sont nettement limitatifs. Nous sommes peu informés sur le type de composante étant extraites par les humains; intuitivement, on serait porté à croire que ces composantes ne sont pas orthogonales, puisque ceci augmenterait inutilement la taille de l'espace psychologique utilisé pour définir les stimulations du monde extérieur (puisque chaque nouvelle composante développée doit nécessairement être orthogonale aux précédentes).

Les modèles non-linéaires (comme ceux à base d'ICA) semblent rendre compte adéquatement de données obtenues en cognition animale (Zhang et Mei, 2003; Hyvärinen, Hurri, et Väyrynen, 2003), et pourraient s'avérer prometteur pour les problématiques décrites. Cependant, à moins de postuler une redondance des composantes dans le système cognitif, ces modèles impliquent une approche localiste de l'information, et par conséquent ne permettent pas une diminution gracieuse des performances. Une solution à ce problème serait

de distribuer l'information sur l'ensemble des unités, ou à tout le moins une partie (Olshausen et Field, 2004; O'Reilly, 1998).

En résumé, l'approche en composantes indépendantes semble la plus performante pour l'extraction des caractéristiques maximisant la reconstruction de l'entrée. Cependant, cette approche implique une architecture contraignante demandant de connaître d'avance le nombre maximal de composantes à extraire. De plus, puisqu'aucun attracteur n'est présent dans le réseau, ces modèles ne tolèrent pas le bruit et l'incomplétude.

#### 2.2.1.2 Règles d'apprentissage compétitives

Dans sa forme la plus stricte, le principe d'apprentissage compétitif (Rumelhart et Zipser, 1986) implique que seule une unité de sortie, l'unité gagnante, peut être activée à la fois (principe *winner-take-all* ou WTA). Ce type d'apprentissage autonome est donc très pratique pour classifier un ensemble de patrons d'entrée à l'intérieur d'un nombre limité de catégories. Puisque l'information y est locale, il est très facile de « lire » le système, contrairement aux perceptrons multi-couches. L'architecture des modèles WTA suit celle d'un réseau de type *feedforward* contenant deux couches d'unités, avec, dans certains cas, présence d'inhibition latérale dans la couche de sortie (comme dans certains réseaux PCA). Cependant, contrairement aux PCA, l'inhibition latérale est symétrique, ce qui assouplit les contraintes architecturales.

La règle de transmission de ce type de réseau est déterminante pour le type de code développé. Elle est exprimée par :

$$G = \underset{i}{Min} \left\| \mathbf{w}_i - \mathbf{x} \right\| \qquad 2.10$$

où $G$ représente l'unité gagnante, $\mathbf{w}_i$ représente le $i$ ème vecteur de la matrice de poids, et $\mathbf{x}$ représente le stimulus d'entrée. Cette règle exprime que seule l'unité liée au vecteur de poids le plus près du vecteur d'entrée sera sélectionnée pour une mise à jour. La règle d'apprentissage pour la mise à jour des poids est exprimée ainsi :

$$\mathbf{w}_i(k) = \mathbf{w}_i(k-1) + \eta(\mathbf{x} - \mathbf{w}_i(k-1)) \qquad 2.11$$

où $k$ représente le numéro d'essai, et $i$ représente l'unité gagnante. Cette règle vise à minimiser la distance entre le vecteur de poids gagnant et le vecteur d'entrée. Ainsi, chaque

entrée devient associée à un vecteur de poids, et par conséquent à une unité de sortie. Dans une situation où les stimuli sont orthogonaux, les poids de connexions finaux seront identiques aux vecteurs de stimuli.

Une famille de modèles compétitifs WTA très populaire est celle des réseaux ART (Grossberg, 1988)[21]. Les modèles ART reposent sur deux idées principales, soit la résonance et la vigilance. Grâce à ces deux principes, ces modèles sont en mesure de répondre au dilemme de stabilité-plasticité (Freeman, 1994). La résonance est un état d'équilibre atteint lorsque la sortie du réseau ne se modifie plus (*i.e.* lorsque le réseau a atteint un attracteur). Contrairement aux réseaux compétitifs décrits précédemment, la transmission se fait par projection (produit interne). Dans le cas du modèle ART1, puisque les stimuli sont binaires et que les poids de connexions représentent les stimuli eux-mêmes, le produit interne revient, une fois une fonction de seuil appliquée, à compter le nombre de valeurs « 1 », et à voir si cette somme correspond à ce qui est toléré par le paramètre de vigilance. Le modèle ART2 utilise un principe équivalent, mais avec des stimuli à valeurs réelles.

Ces réseaux (ART1, ART2) apprennent seulement lorsque l'état d'équilibre (ou de résonance) est atteint. Pour déterminer si la reconstruction est suffisante, une mesure de similarité entre l'entrée et sa reconstruction par la sortie est comparée à un critère, le paramètre de vigilance. Ce critère ne requiert pas des reconstructions parfaites, mais plutôt des reconstructions suffisamment semblables. La valeur donnée préalablement au paramètre de vigilance est extrêmement importante et variera en fonction de la tâche. Une valeur élevée pour ce paramètre mène à la production de nombreuses mémoires hautement détaillées (exemplaires), tandis qu'une valeur basse produit des mémoires plus générales (prototypiques). Si aucune des unités présente dans le réseau ne permet à ce dernier d'atteindre un état de résonance, le réseau ajoutera une nouvelle unité et les poids de cette nouvelle unité correspondront tout simplement au stimulus lui-même (voir Hélie, Chartier, et Proulx, 2006, pour une description formelle du type de représentation développée par le réseau ART1).

[21] Les réseaux ART n'utilisent pas la règle de transmission proposée par Rumelhart et Zipser (1986).

La famille de réseaux ART possède aussi une version supervisée. Le réseau ARTMAP (Carpenter, Grossberg, et Reynolds, 1991) joint deux réseaux ART légèrement modifiés en une structure unique supervisée, où une première unité compétitive utilise les données d'entrée, et une autre unité utilise la sortie désirée du réseau. Ces deux unités séparées sont utilisées dans le but d'ajuster le paramètre de vigilance pour réussir la classification adéquate du patron d'entrée.

Il semble donc que les réseaux ART puissent répondre à la grande majorité des exigences posées. Ils peuvent effectuer de l'extraction de caractéristiques, produisent des représentations extrêmement comprimées (en fait, en une seule dimension), peuvent modéliser des mémoires d'exemplaires et de catégories, et peuvent effectuer du traitement autonome et supervisé.

Malheureusement, certains principes sont inadéquats pour le projet. Pour débuter, l'utilisation du principe WTA rend (tout comme pour les réseaux à base de PCA) le système extrêmement fragile, puisque la perte d'une unité dans le réseau correspond automatiquement à la perte d'un exemplaire ou d'une catégorie. Aussi, ce type de codage est en fait une représentation localiste, qui, au niveau biologique serait l'équivalent d'un neurone extrêmement spécifique, d'une « cellule grand-mère ». Même si la présence de ce type de cellule dans le cortex a été montrée, il ne semble pas que les représentations localistes soient un principe général au niveau neurologique (Gross, 2002). O'Reilly (1998) rapporte d'ailleurs que le principe des représentations distribuées a plus fréquemment été validé suite à des enregistrements électrophysiologiques.

De façon plus importante, les réseaux compétitifs de type WTA montrent un autre défaut de taille : ils ne sont pas sensibles à la fréquence des exemplaires vus. En effet, Hélie, Chartier et Proulx (2006) ont comparé des représentants de deux classes de modèles non-supervisés, soit les mémoires autoassociatives récurrentes (RAM), à base d'apprentissage hebbien, et les modèles compétitifs de type WTA. Ils ont montré que les réseaux compétitifs (en particulier, les réseaux de Rumelhart et Zipser, ainsi que le réseau ART1), contrairement aux RAM, ne pouvaient rendre compte de différentes distributions de fréquences de présentations d'exemplaires. De nombreux auteurs (entre autres : Erickson et Kruschle, 1998;

Kruschke, 1996; Nosofsky, 1988) ont montré que cette fréquence avait une influence sur les jugements catégoriels chez l'humain. Ceci constitue donc une faille de taille.

### 2.2.2 Réseaux supervisés

#### 2.2.2.1 Supervision interne

Certains réseaux de neurones, les perceptrons multicouches plus précisément, peuvent nécessiter une supervision « interne ». Ce type de supervision est lié à la présence d'un tuteur faisant partie intégrante du réseau, sans jamais être explicitement représenté dans l'architecture. Ce tuteur, connaissant l'activation finale désirée pour la couche de sortie, permet au réseau d'ajuster de graduellement ajuster ses connexions pour correctement reproduire les correspondances entre les entrées et les sorties. On utilise généralement un algorithme de rétropropagation de l'erreur (Rumelhart, Hinton et Williams, 1986) pour ajuster les poids de connexions suite à une rétroaction reçue par le réseau. Ceci correspond en fait à un traitement local menant à l'approximation des paramètres d'une fonction globale (Hornik, Stinchcombe et White, 1989).

Le réseau suit la règle *delta* (Freeman, 1994) pour ajuster ses poids de connexion. La modification des poids de connexion peut être exprimée ainsi pour la matrice liant les unités cachées aux unités de sortie:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta\delta(\mathbf{y}^c)^T \qquad 2.12$$

et ainsi, pour la matrice liant les unités d'entrée aux unités cachées:

$$\mathbf{W}^c(k+1) = \mathbf{W}^c(k) + \eta(\mathbf{y}^c \times (1-\mathbf{y}^c) \times \mathbf{W}^T\delta)\mathbf{x}^T \qquad 2.13$$

tel que:

$$\delta = (\mathbf{d}-\mathbf{y}) \times \mathbf{y} \times (1-\mathbf{y}) \qquad 2.14$$

Ici, $\mathbf{W}$ représente la matrice de poids de connexion entre les unités cachées et les unités de sortie, $\mathbf{W}^c$ représente la matrice de poids entre les unités d'entrée et les unités cachées, $\mathbf{x}$ représente le vecteur d'entrée, $\mathbf{y}$ représente la sortie du réseau, $\mathbf{y}^c$ représente la sortie de la couche cachée, $\mathbf{d}$ représente la sortie désirée, et $\eta$ représente un paramètre d'apprentissage.

Les perceptrons multicouches à base de rétropropagation ont la capacité d'extraire les caractéristiques par la couche d'unités cachées (Negnevitsky, 2002). Malheureusement, malgré les performances spectaculaires auquel il mène, ce type de processus n'est pas considéré plausible biologiquement (Stork, 1989[22]). Il existe des neurones effectuant la rétropropagation de potentiels d'action dans le cerveau, et causant le changement des connexions présynaptiques (Bogacz, Brown et Giraud-Carrier, 2000). Cependant, il n'existe pas au niveau neurologique un tel mécanisme permettant de propager l'erreur vers l'arrière. De façon plus importante, le tuteur représente en fait un homoncule implicite au réseau; ainsi, une composante du réseau connaît la réponse avant l'analyse dynamique, ce qui rend caduque la recherche d'une telle réponse.

Une classe de réseaux connexe aux réseaux à rétropropagation, et qui partage certains liens avec les modèles à extraction de composantes (PCA, nPCA, ICA) est celle des auto-encodeurs (*e.g.*, Hinton et Salakhutdinov, 2006). Les auto-encodeurs doivent par définition contenir une couche d'entrée et une couche de sortie de dimensionnalité égale, et au moins une couche d'unités cachées dont la dimensionnalité est inférieure. Ces réseaux sont utilisés à prime abord pour des applications d'apprentissage machine, et sont donc peu utilisées pour des problématiques d'apprentissage cognitif. À la base, les auto-encodeurs apprennent à comprimer les patrons d'entrée et extraire un code réduit (ou vocabulaire de composantes), tel que requis pour ce projet.

Malgré le caractère autonome de certains auto-encodeurs (grâce à l'utilisation de machines de Boltzmann : Freeman, 1994), la plupart de ces réseaux, parce qu'ils visent l'apprentissage machine, ont des paramètres réglés à l'aide d'algorithmes de rétropropagation. Aussi, la rétropropagation doit être utilisée pour les problématiques de catégorisation et de développement de mémoires d'exemplaires. Finalement, par leur architecture *feedforward,* les auto-encodeurs perdent des capacités de tolérance au bruit et à l'incomplétude.

[22] En fait, à l'origine, Rumelhart, Hinton et Williams (1986) ont eux-mêmes admis que l'algorithme de rétropropagation n'était pas un modèle plausible de l'apprentissage au niveau cérébral.

2.2.2.2 Supervision externe

Une généralisation directe des modèles de type RAM est la classe des mémoires associatives bidirectionnelles (BAM; Kosko, 1988). Ces mémoires peuvent associer entre eux deux vecteurs de taille égale ou différente (représentant, par exemple, une entrée visuelle avec une représentation comprimée). Ces réseaux possèdent l'avantage de constituer des mémoires autoassociatives et hétéroassociatives, et par conséquent peuvent couvrir plusieurs cas liés à l'apprentissage et la catégorisation perceptuels. Par leur architecture, elles présentent également l'avantage de pouvoir effectuer de l'apprentissage supervisé[23] ou non-supervisé.

À l'instar du modèle d'Hopfield (1982), dans la BAM standard de Kosko (1988), l'apprentissage est accompli grâce à l'utilisation d'une règle hors-ligne strictement hebbienne, suivant l'équation suivante :

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^{\mathbf{T}} \qquad 2.15$$

Dans cette expression, les matrices **X** et **Y** représentent les ensembles de paires de vecteurs bipolaires qui doivent être associés, et **W** représente la matrice de poids qui, en fait, se veut une mesure de corrélation entre l'entrée et la sortie du réseau. Le modèle utilise une fonction de transmission non-linéaire récurrente pour permettre la filtration des différents patrons durant une tâche de rappel. Tout comme le modèle d'Hopfield, la BAM de Kosko utilise une fonction de transmission de type *signum* pour effectuer le rappel à partir de patrons bruités. La fonction de transmission non-linéaire généralement utilisée dans les réseaux de type BAM suit les équations suivantes :

$$\mathbf{y}(t+1) = \mathrm{sgn}(\mathbf{W}\mathbf{x}(t)) \qquad 2.16$$

et

$$\mathbf{x}(t+1) = \mathrm{sgn}(\mathbf{W}^T\mathbf{y}(t))$$

où **y**(t) et **x**(t) représentent la paire de vecteurs choisis lors de l'essai *t*, **W** représente la matrice de poids, et *sgn* est la fonction *signum*.

[23] L'apprentissage supervisé peut être réalisé par l'association d'un ensemble de stimuli avec un ensemble de réponses prédéterminées.

La BAM de Kosko, malgré son élégance, présente plusieurs caractéristiques indésirables pour la modélisation de processus humains. Par exemple, l'apprentissage n'y est pas effectué en ligne, le réseau produit beaucoup d'attracteurs nuisibles, et sa capacité de stockage est limitée. Au cours des dernières années, plusieurs améliorations et ajouts au principe de base de la BAM ont été proposés (entre autres : Leung, 1994; Wang, 1996). Alors que la plupart des améliorations visaient à augmenter la capacité de stockage, réduire la quantité d'attracteurs fortuits, et améliorer la performance du modèle, toutes ces méthodes sont associées à un coût énorme en ce qui a trait à la complexification de l'architecture et de la procédure de fonctionnement (Chartier et Boukadoum, 2006).

Chartier et Boukadoum (2006) ont proposé d'appliquer les avantages du modèle NDRAM à la classe plus générale des BAM. Pour ce faire, ils ont proposé une architecture bidirectionnelle hétéroassociative (BHM), contenant deux matrices de poids de connexions distinctes, soit **W** et **V**. L'utilisation de deux matrices séparées permet d'augmenter la plausibilité du modèle, en permettant la création et la mise à jour de connexions asymétriques entre les deux couches d'unités (Hassoun, 1989). Ils ont appliqué au processus les règles d'apprentissage et de transmission utilisées par Chartier et Proulx (2005), de sorte que la règle d'apprentissage est maintenant exprimée par :

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(\mathbf{y}(0) - \mathbf{y}(t))(\mathbf{x}(0) - \mathbf{x}(t))^T \qquad 2.17$$

et

$$\mathbf{V}(k+1) = \mathbf{V}(k) + \eta(\mathbf{x}(0) - \mathbf{x}(t))(\mathbf{y}(0) - \mathbf{y}(t))^T \qquad 2.18$$

La règle de transmission est maintenant exprimée par :

$$\forall i,...,N, \mathbf{y}_i(t+1) = \begin{cases} 1, \ Si \ \mathbf{W}\mathbf{x}_i(t) > 1 \\ -1, \ Si \ \mathbf{W}\mathbf{x}_i(t) < -1 \\ (\delta+1)\mathbf{W}\mathbf{x}_i(t) - \delta(\mathbf{W}\mathbf{x})_i^3(t), \ Sinon \end{cases} \qquad 2.19$$

et

$$\forall i,...,N, \mathbf{x}_i(t+1) = \begin{cases} 1, \ Si \ \mathbf{V}\mathbf{y}_i(t) > 1 \\ -1, \ Si \ \mathbf{V}\mathbf{y}_i(t) < -1 \\ (\delta+1)\mathbf{V}\mathbf{y}_i(t) - \delta(\mathbf{V}\mathbf{y})_i^3(t), \ Sinon \end{cases} \qquad 2.20$$

Ces modifications ont permis, tout comme avec NDRAM, l'apprentissage en ligne, le traitement de vecteurs à valeurs continues, une performance de rappel supérieure, et une réduction du nombre d'attracteurs nuisibles.

Une fois de plus, malgré l'élégance du modèle, la BHM ne répond pas tout à fait aux propriétés désirables du modèle qui sera proposé. Le modèle peut associer deux vecteurs de dimensionnalité différente, mais à condition que ces deux vecteurs soient disponibles dès le début de la tâche. Aussi, le réseau performe bien en autant que l'information soit déjà nettoyée de son bruit, et que les appartenances aux différentes classes sont bien définies. Pour ce projet, le but est plutôt que le réseau en vienne à associer un patron d'entrée avec une version économique basée sur un ensemble de caractéristiques développées de façon autonome. Plusieurs caractéristiques intéressantes du modèle pourront cependant être conservées, telles les matrices de connexions distinctes, et l'association bidirectionnelle.

## 2.3 Conclusion

En résumé, il semble donc qu'aucune des classes de base de réseaux de neurones présentées ne répond à toutes les hypothèses de travail établies. Du côté des modèles autonomes, les RAM peuvent mémoriser des exemplaires, généraliser autour de prototypes, et sont tolérantes au bruit et à l'incomplétude Cependant, elles ne peuvent former des catégories et comprimer l'information. Les réseaux à base d'extraction de composantes réussissent à extraire les caractéristiques et comprimer les représentations, mais leur architecture est limitative, et empêche la résistance au bruit et à l'incomplétude. Aussi, le traitement mène à des représentations localistes (O'Reilly, 1998). Les réseaux compétitifs de type WTA peuvent effectuer la réduction dimensionnelle, l'extraction de composantes et la catégorisation à proprement dire. Cependant, les représentations y sont aussi locales, et ces réseaux ne reproduisent pas les biais environnementaux (Hélie, Chartier et Proulx, 2006).

Du côté des modèles supervisés, les réseaux basés sur la rétropropagation sont à proscrire, puisqu'ils ne sont pas plausibles psychologiquement. Les auto-encodeurs, que l'on pourrait qualifier de « semi-supervisés », réduisent le code et extraient des composantes sans limitations d'orthogonalité, mais une fois de plus, leur architecture unidirectionnelle les rend

intolérants au bruit et à l'incomplétude. Finalement, les BAM constituent d'excellents candidats pour l'association entre un stimulus et sa représentation comprimée, ou encore entre une compression et une représentation catégorielle. Cependant, le fait que ces réseaux nécessitent deux entrées simultanées à l'apprentissage fait qu'ils ne peuvent être utilisés directement.

Il semble donc que la solution devra donc passer par une mise en commun de deux types de réseau. La combinaison la plus parcimonieuse serait celle des RAM et des BAM, puisque le premier type de réseau constitue une spécification du second. Ainsi, en utilisant ces deux classes de réseau, on réduirait le nombre de postulats de base nécessaires à la définition de l'approche. Dans le prochain chapitre, un intermédiaire entre ces deux classes de réseaux sera proposé.

# CHAPITRE III

# FEBAM : UNE MÉMOIRE ASSOCIATIVE BIDIRECTIONNELLE EXTRACTRICE DE CARACTÉRISTIQUES

Dans le chapitre précédent, il a été conclu qu'aucune des classes de modèles présentées ne pouvait complètement satisfaire les exigences cognitives établies pour le modèle perceptivo-cognitif désiré. Cependant, plusieurs classes possédaient une partie de la réponse, un élément désirable de la solution. En intégrant les caractéristiques de ces quelques classes de réseaux, il sera possible de proposer un premier modèle effectuant des processus d'extraction de caractéristiques et de réduction dimensionnelle, ainsi que de développement d'exemplaires (identification) et de catégorisation, et ce, de façon autonome.

Les réseaux à extraction de composantes (PCA, nPCA, ICA) ont pour but la réduction dimensionnelle. Cependant, il n'est pas encore clair que les composantes utilisées au niveau perceptuel chez l'humain suivent des contraintes aussi rigides d'orthogonalité ou d'indépendance. Une chose est toutefois sûre : réduction dimensionnelle et reconstruction de stimuli sont définis, dans ces réseaux, comme un processus unitaire. Par leur architecture unidirectionnelle, ces réseaux ne sont donc pas tolérants au bruit et à l'incomplétude, comme le sont les RAM. Cela dit, le critère de réussite des réseaux à extraction de composantes (*i.e.* la fonction d'erreur), qui vise une reconstruction de qualité maximale de l'entrée par un apprentissage autonome, mérite d'être conservé sous une forme ou une autre.

La classe des auto-encodeurs peut être vue comme une généralisation des réseaux à extraction de composantes. Alors que ces derniers utilisent une « couche de compression » (la couche de sortie), les auto-encodeurs en utilisent généralement plusieurs. Ces derniers nécessitent donc plusieurs couches d'unités cachés, ce qui est difficile d'adapter en apprentissage continu. Par conséquent, on doit segmenter l'apprentissage en entrainant les

couches deux à deux et/ou en utilisant un algorithme de descente de gradient (tel que la rétropropagation). Dans un cas comme dans l'autre, la réduction dimensionnelle est coûteuse sur le plan de la consistance interne (ainsi que sur le plan de la plausibilité biologique). Cela dit, ces algorithmes d'apprentissage machine sont en mesure de définir un code réduit adapté à la reconstruction des entrées[24], et ce, sans contraintes rigides quant à l'orthogonalité ou l'indépendance des composantes du code. Ce type de modèle pourrait donc servir à réaliser les tâches précédemment décrites s'ils peuvent être adaptés à la modélisation cognitive (en incluant les contraintes décrites précédemment). Plus précisément, l'ajout d'une boucle de récurrence permettrait d'ajouter une tolérance au rappel bruité et incomplet.

La classe des mémoires autoassociatives récurrentes (RAM) permet d'effectuer la différentiation des stimuli, au sens entendu par Gibson (1969). En effet, le but principal de ce type de réseau est la reconstruction de l'entrée (ou récupération du vecteur d'origine), qui constitue une forme de différentiation « extrême ». Pour ce faire, il faut faire correspondre les attracteurs du système aux stimuli à différencier. Ainsi, les bassins d'attraction permettront à des stimuli similaires (qui se situent à l'intérieur d'un rayon d'attraction donné) de converger vers les attracteurs voulus. Par conséquent, ils sont en mesure d'effectuer une reconstruction (ou récupération en mémoire) parfaite à partir de rappels avec ajout de bruit gaussien (rappel bruité) ou inversion de polarité de pixels (rappel incomplet). Cependant, les RAM ne peuvent produire de représentations cognitivement économiques. Chaque unité étant généralement connectée à toutes les autres unités du réseau[25], il s'ensuit que le nombre de connexions nécessaires pour le rappel parfait est de l'ordre de $u^2$, où $u$ représente le nombre d'unités du réseau. Pour effectuer la compression tout en conservant les caractéristiques des RAM, il faudrait ajouter une couche intermédiaire contenant un nombre inférieur d'unités (principe inspiré des auto-encodeurs) et effectuer un apprentissage bidirectionnel (un principe

[24] C'est-à-dire que les vecteurs de l'ensemble original peuvent être redéfinis à l'aide d'un code à dimensionalité réduite, et qu'à l'aide de ce code, une perte d'information minime (ou jugée acceptable) est produite par le processus.

[25] Sauf pour certains réseaux de type Hopfield, où les autoconnexions sont absentes pour des raisons de performance.

computationnel notamment supporté par O'Reilly (1998), qui constituerait également selon lui un principe de modélisation d'inspiration biologique).[26]

Le réseau devra associer une entrée avec la compression correspondante qu'il aura créée. L'association entre deux vecteurs se fait généralement par le biais de mémoires associatives bidirectionnelles (BAM). Or, la valeur des unités de compression est inconnue d'avance. Autrement dit, puisque le problème de l'autonomie exclut un superviseur, les modèles BAM ne peuvent directement résoudre le problème posé. De plus, l'utilisation d'une seule matrice de connexions (et de sa transposée) pour transférer l'information d'une couche à l'autre (tel qu'avec les BAM et autoencodeurs n'est pas justifiée sur le plan de la tâche. En effet, ces deux matrices ont des buts différents : la première matrice doit créer une mémoire de correspondances entre les stimuli et leur version comprimée, alors que la deuxième couche doit créer une mémoire de composantes permettant le passage d'une représentation comprimée à une reconstruction (qui selon les circonstances, sera parfaite ou différenciable). Quoique complémentaires, ces solutions ne sont pas nécessairement symétriques. Par conséquent, un principe d'asymétrie proposé par Hassoun (1989) sera privilégié; ainsi, on devra utiliser deux matrices séparées dans le réseau.

Dans le présent chapitre, il sera proposé que pour conserver les caractéristiques désirées de toutes ces classes de modèles, il suffit d'ajouter une couche d'unités dites de *compression* au modèle autoassociatif récurrent qu'est NDRAM. Ceci peut également être vu comme une modification au BHM de Chartier et Boukadoum (2006), où l'on ignorerait au préalable le contenu de l'une des couches. Ainsi, le modèle résultant devient un auto-encodeur auto-supervisé à connexions asymétriques, qui devrait être tolérant au bruit et à l'incomplétude des entrées au rappel.

[26] Une économie au niveau de la taille des représentations sera effectuée aussitôt que le nombre d'unités de la couche de intermédiaire est inférieur à $u$. Si l'on considère le nombre de connexions du réseau, puisqu'une couche supplémentaire d'unités (et de connexions) est ajoutée, il faut que le nombre d'unités de compression soit inférieur à $u/2$ afin qu'une économie soit réalisée.

## 3.1 Description du modèle

Comme tout réseau de neurones artificiels, il est possible de décrire FEBAM (Feature-Extracting Bidirectional Associative Memory) en présentant son architecture, sa règle de transmission et sa règle d'apprentissage.

### 3.1.1 Architecture

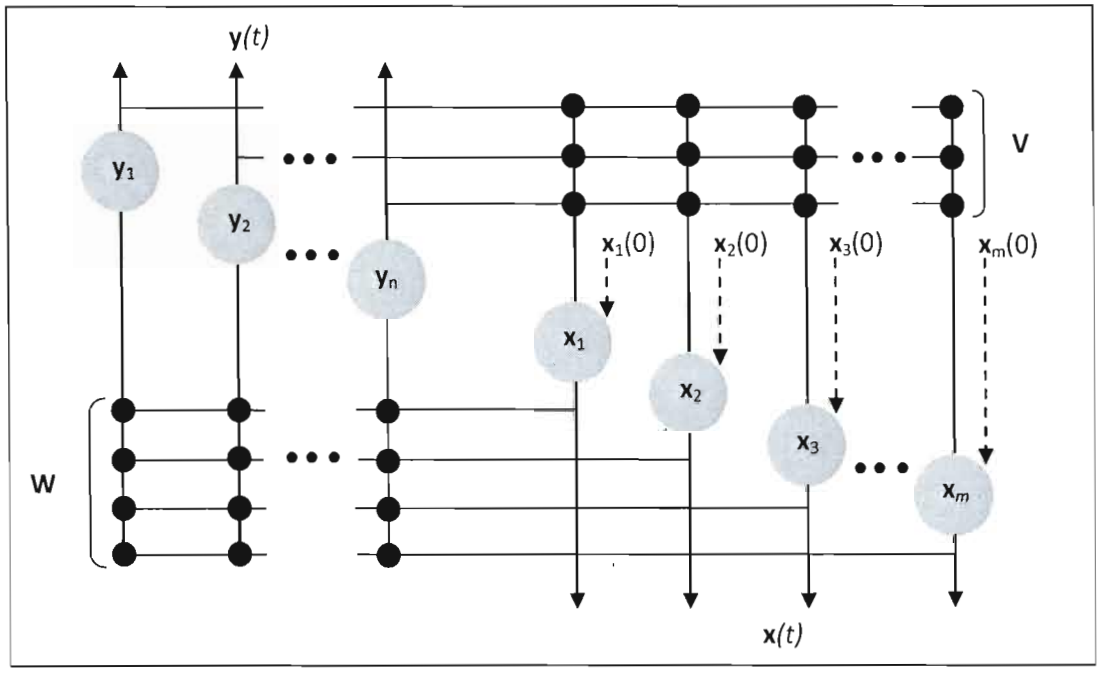


**Figure 3.1** Architecture neuronale du modèle FEBAM. Le réseau contient deux couches d'unités, et chaque unité de la couche **x** est connectée à chaque unité de la couche **y**. Il n'y aucune connexion entre unités d'une même couche. Les lignes pointillées indiquent le point d'entrée du réseau. Ainsi, les stimuli (entrée initiale **x**(0)) seront « présentés » à la couche **x**.

Le modèle FEBAM (Chartier, Giguère, Renaud, Lina, et Proulx, 2007; Giguère, Chartier, Proulx, et Lina; 2007a, 2007b), tel que déjà mentionné, constitue un compromis entre une mémoire autoassociative récurrente, en l'occurrence NDRAM (Chartier et Proulx, 2005), et une mémoire hétéroassociative bidirectionnelle (BHM : Chartier et Boukadoum, 2006), et ceci se reflète entre autres au niveau de l'architecture (Figure 3.1). Le réseau consiste en deux réseaux de neurones de type Hopfield, doublement interconnectés, et

l'activation est bidirectionnelle (ascendante/descendante). Comme pour une BAM standard, chaque couche d'unité peut servir de « tuteur interne » pour l'autre couche.

Le modèle contient donc deux couches, soit **x** et **y**. La couche **x** sert d'entrée initiale au réseau, ainsi que de couche de reconstruction en vue d'une comparaison avec l'entrée. La taille de cette couche (en nombre d'unités) est égale au nombre de pixels des images d'origine. La couche d'unités **y** sert à compresser l'information, *i.e.* appliquer et mémoriser une réduction dimensionnelle. Le niveau de compression est prédéfini, et dépend directement du nombre d'unités de la couche **y** (aussi nommées *unités de compression*). Lorsqu'il n'y a qu'une seule unité dans la couche **y**, la compression est maximale; lorsque le nombre d'unités **y** est minimalement égal à celui de la couche **x**, aucune compression n'est effectuée.

Tels certains modèles de mémoire associatives bidirectionnelles (*e.g.*, Chartier et Boukadoum, 2006); deux matrices de poids de connexion **V** et **W** sont utilisées pour représenter la possibilité de connexions asymétriques, formant ainsi deux mémoires distinctes. Le contenu de ces mémoires n'est pas totalement indépendant (compte tenu du critère de réussite qui sera utilisé), mais la mise à jour se fait de façon séparée, sans usage de rétropropagation. Dans le cas de FEBAM, les deux matrices utilisées auront ici deux rôles bien précis, dépassant la simple asymétrie.

La matrice **W**, reliant la couche **x** à la couche **y,** sert à déterminer la fonction permettant de transformer les intrants en leur version comprimée (matrice de compression). La matrice **V**, joignant la couche **y** à la couche **x**, permet, tel un auto-encodeur, de reconstruire le stimulus à l'aide des filtres (ou composantes) appris (matrice de reconstruction). Cette matrice contient donc le vocabulaire de composantes (les « blocs de construction ») extraites par le réseau.

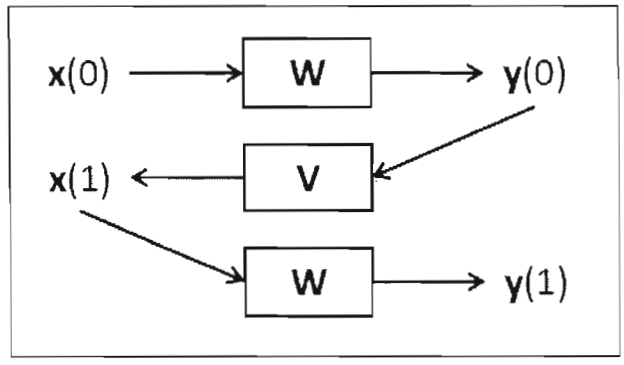


**Figure 3.2** Schéma illustratif du processus itératif (ou *cycle*) réalisé par le réseau avant chacune des mises à jour des matrices de poids de connexion. Dans la présente thèse, le nombre de cycles réalisés avant la mise à jour des poids de connexion sera toujours égal à 1.

Ici, la différence principale entre le BHM et FEBAM est l'absence d'une entrée « externe » dans la couche **y**. En effet, puisqu'une mémoire bidirectionnelle sert à associer deux ensembles de vecteurs connus, pour chaque stimulus **x**, il doit aussi exister un stimulus **y**. Dans le cas de FEBAM, il n'y aucune entrée initiale pour la couche **y**. Le contenu de cette couche sera créé lors du processus initial de compression, au cours duquel on associera l'entrée provenant de l'environnement avec sa propre version comprimée. Cette dernière pourra ensuite être utilisée, si désiré, comme représentation vectorielle d'entrée pour un autre modèle.

La Figure 3.2 illustre le processus suivi par FEBAM lors de la présentation d'un stimulus à l'apprentissage. Pour débuter, FEBAM transforme l'entrée à dimensionnalité complète **x**(0) en une version réduite ou comprimée d'elle-même **y**(0), à l'aide du produit entre la matrice **W** et le vecteur d'entrée. Ensuite, cette version comprimée passe à travers la matrice **V**, dans le but de produire une reconstruction finale **x**(1) la plus fidèle possible au stimulus original à l'aide du« vocabulaire » de composantes développées de façon autonome. Cette reconstruction passe une deuxième fois par la matrice **W**, produisant la compression finale **y**(1)[27].

[27] Ici, le double passage dans la matrice **W** est nécessaire pour faire la mise à jour des poids de connexion, basée sur une différence temporelle.

### 3.1.2 Règle de transmission

La règle de transmission du réseau FEBAM a été proposée par Chartier et ses collègues pour NDRAM et le BHM, pour des raisons plus techniques que cognitives. En effet, l'utilisation des fonctions habituelles dans les réseaux récurrents (sigmoïde et ou tangente hyperbolique) ne permet pas la création d'attracteurs positionnés à l'une des surfaces de l'hypercube, car ces fonctions sont asymptotiques. Dans le cas des BAM, la fonction généralement utilisée est une fonction de type *signum* (voir Kosko, 1988), qui, à l'opposé, ne permet pas la production d'attracteurs autres que binaires ou bipolaires dans l'espace. La fonction proposée par Chartier et al. (Figure 3.3) constitue un compromis, puisqu'elle permet la production d'attracteurs en toute position de l'espace. Ainsi, elle peut mener à la création d'attracteurs à valeurs continues (*i.e.* en tons de gris; Chartier et Proulx (2005) fournissent un exemple de ce comportement avec NDRAM).

Cette fonction *f*, lorsqu'utilisée dans le cadre de FEBAM, est exprimée par les équations suivantes :

$$\forall i,...,N, \mathbf{y}_i(t) = \begin{cases} 1, \ Si \ \mathbf{W}\mathbf{x}_i(t) > 1 \\ -1, \ Si \ \mathbf{W}\mathbf{x}_i(t) < -1 \\ (\delta+1)\mathbf{W}\mathbf{x}_i(t) - \delta(\mathbf{W}\mathbf{x})_i^3(t), \ Sinon \end{cases} \quad 3.1$$

$$\forall i,...,M, \mathbf{x}_i(t+1) = \begin{cases} 1, \ Si \ \mathbf{V}\mathbf{y}_i(t) > 1 \\ -1, \ Si \ \mathbf{V}\mathbf{y}_i(t) < -1 \\ (\delta+1)\mathbf{V}\mathbf{y}_i(t) - \delta(\mathbf{V}\mathbf{y})_i^3(t), \ Sinon \end{cases} \quad 3.2$$

où $N$ représente le nombre d'unités dans la couche de compression $\mathbf{y}$, $M$ représente le nombre d'unités dans la couche d'entrée $\mathbf{x}$, $i$ représente l'indice de l'élément du vecteur respectif, $\mathbf{y}(t)$ et $\mathbf{x}(t)$ représentent les contenus des couches d'unités au temps $t$, $\mathbf{W}$ représente la matrice de compression et $\mathbf{V}$ représente la matrice de reconstruction. $\delta$ est un paramètre général de transmission, qui selon Chartier et Proulx (2005), doit être préalablement fixé entre 0 et 0.5 pour assurer des comportements d'attracteurs stables (par opposition à cycliques ou chaotiques; voir également Hélie (2008)).

Dans le cas présent, la fonction de transmission nécessitera l'initialisation aléatoire des poids de connexion dans les deux matrices du réseau à l'aide de valeurs près de zéro. Pour

NDRAM et le BHM, les matrices de poids sont initialisées à l'aide de valeurs nulles (même s'il a été montré que l'initialisation aléatoire n'a aucun impact sur la convergence et ne fait que ralentir le processus d'apprentissage). L'apprentissage est en fait impossible sous cette condition dans FEBAM, puisque le résultat d'un passage dans la fonction mènerait toujours au même résultat, soit zéro. L'impact de cette modification à la procédure devra être testé, pour s'assurer que la réussite d'une simulation ne soit pas due à une quelconque initialisation précise des poids de connexion.

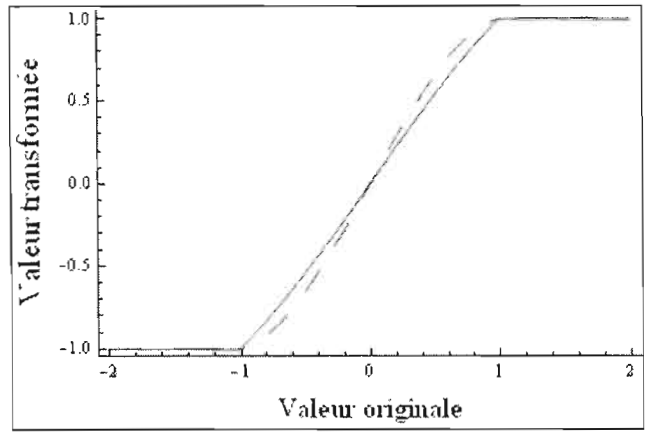


**Figure 3.3** Illustration de la fonction de transmission utilisée pour FEBAM, lorsque le paramètre général de transmission $\delta$ est égal à 0.4 (ligne pointillée) et 0.1 (ligne pleine). Cette dernière valeur sera utilisée pour toutes les simulations présentées dans cette thèse. Bien qu'il n'existe aucune justification cognitive pour ce choix, ce dernier respecte le critère de stabilité défini à l'équation 3.5, et a été utilisé par Chartier et Proulx (2005). Pour éviter un quelconque effet différentiel d'une simulation à l'autre, ce paramètre ne sera donc pas varié.

### 3.1.3 Règle d'apprentissage

L'apprentissage dans FEBAM suit un principe d'association hebbienne basée sur les différences temporelles (Chartier et Boukadoum, 2006; Chartier et Proulx, 2005; Kosko, 1990; Oja, 1989; Sutton, 1988), et peut être formellement exprimé à l'aide des équations suivantes :

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(\mathbf{y}(0) - \mathbf{y}(t))(\mathbf{x}(0) + \mathbf{x}(t))^{\mathrm{T}} \qquad 3.3$$

$$\mathbf{V}(k+1) = \mathbf{V}(k) + \eta(\mathbf{x}(0) - \mathbf{x}(t))(\mathbf{y}(0) + \mathbf{y}(t))^{\mathrm{T}} \qquad 3.4$$

où $\mathbf{W}(k)$ et $\mathbf{V}(k)$ représentent les contenus des matrices de poids de connexion lors de l'essai d'apprentissage $k$, $\mathbf{x}(0)$ représente l'entrée initiale du réseau, $\mathbf{y}(0)$ représente la compression

initiale, $\mathbf{y}(t)$ et $\mathbf{x}(t)$ représentent les vecteurs d'état finaux après $t$ itérations dans le réseau, et $\eta$ est un paramètre général d'apprentissage. Cette règle est somme toute plutôt simple, et constitue une généralisation du principe de corrélation hebbienne/anti-hebbienne dans sa forme autoassociative. L'apprentissage hebbien, tel que déjà mentionné au chapitre 2, est généralement utilisé dans les réseaux autonomes parce qu'il est considéré biologiquement plausible (O'Reilly, 1998). Le principe hebbien/anti-hebbien, aussi utilisé par Bégin et Proulx (1996), est utilisé parce qu'il constitue une solution locale au problème de croissance infinie des poids de connexions (tel qu'observé dans le BSB).

Suite à une analyse théorique de Chartier et Proulx (2005; Chartier et Boukadoum, 2006), il a été déterminé que pour permettre une convergence des poids, le paramètre d'apprentissage lié à cette règle doit être fixé selon la condition suivante :

$$\eta < \frac{1}{2(1-2\delta)Max\left[N,M\right]},\ \delta \neq 1/2 \qquad 3.5$$

Cette condition doit être respectée pour produire la stabilité, caractéristique importante de l'apprentissage par un système cognitif.

Les équations 3.3 et 3.4 montrent que les poids de connexion ne peuvent converger que si la compression finale $\mathbf{y}(t)$ est égale à la compression initiale $\mathbf{y}(0)$ (convergence des poids de la matrice $\mathbf{W}$), ou lorsque la reconstruction finale $\mathbf{x}(t)$ est identique à l'entrée initiale $\mathbf{x}(0)$ (convergence des poids de la matrice $\mathbf{V}$). Optimalement, si le réseau extrait des caractéristiques permettant de parfaitement reconstruire l'entrée initiale, la différence entre $\mathbf{x}(t)$ et $\mathbf{x}(0)$ sera nulle. Le critère d'arrêt et de réussite du réseau dépendra de la tâche à accomplir : pour une tâche de développement d'exemplaires optimaux, la reconstruction sera prise en compte, alors que pour une tâche de catégorisation non-supervisée, la stabilité de la représentation comprimée sera recherchée.

Dans le cadre de cette thèse, le nombre d'itérations $t$ avant la mise à jour des poids de connexions sera toujours égal à 1. L'utilisation d'un nombre de cycles plus élevé ne change pas qualitativement la performance du réseau. Ce choix cadre avec une optique selon laquelle une mémoire de chaque entrée perceptuelle est toujours disponible au système lorsque se produit la comparaison ultime entre la reconstruction finale et l'entrée originale. Ceci serait

cognitivement supporté par le fait que l'humain possède une mémoire dite *iconique* (Goldstein, 2008; Sperling, 1960), non-interprétée, et d'une durée d'environ 250 millisecondes (en ce qui a trait à la perception visuelle). Ainsi, le traitement doit être court si le système veut profiter de la présence de l'entrée originale dans la mémoire pour effectuer une comparaison (d'où le choix de $t = 1$).

Ainsi, compte tenu du critère de réussite recherché, FEBAM peut être vu comme émulant un réseau d'analyse en composantes non-linéaire tentant de minimiser l'une des fonctions d'erreurs suivantes:

$$E = \|\mathbf{x}(1) - \mathbf{x}(0)\|^2 \tag{3.6}$$

ou

$$E = \|\mathbf{y}(1) - \mathbf{y}(0)\|^2 \tag{3.7}$$

Si l'on se concentre sur le critère de reconstruction (Équation 3.6), on voit ici que lors des simulations, FEBAM tentera de minimiser la différence entre l'entrée initiale et la reconstruction finale. Étant donné l'utilisation de deux matrices distinctes, et d'une fonction de transmission non-linéaire dans le réseau, l'équation 3.6 équivaut à :

$$E = \|f(\mathbf{V}\mathbf{y}(0)) - \mathbf{x}(0)\|^2 = \|f(\mathbf{V}f(\mathbf{W}\mathbf{x}(0)) - \mathbf{x}(0)\|^2 \tag{3.8}$$

ce qui est directement comparable à l'équation de la fonction d'erreur pour l'analyse en composantes principales non-linéaire (Équation 2.9), où l'entrée du réseau passe par la matrice de poids et par sa transposée avant d'être comparée à sa version initiale. Dans les deux cas, une reconstruction de l'entrée est directement comparée avec l'entrée originale dans le but de déterminer le caractère adéquat du codage développé par le réseau.

## 3.2 Les utilisations potentielles de FEBAM

Les deux matrices séparées de FEBAM permettront l'utilisation du réseau avec différents objectifs perceptivo-cognitifs différentes : simple extraction de caractéristiques, développement d'une mémoire d'exemplaires minimale (prédifférentiation) ou optimale, catégorisation et identification non-supervisée (et par extension: reconnaissance et discrimination). Le postulat de base de l'approche est que le système cognitif, lorsqu'il est

non-supervisé et ne vise aucun but précis, effectue un simple apprentissage perceptuel, visant à terme l'optimalité de la reconstruction (différentiation parfaite) par la création d'un vocabulaire de composantes. Ainsi, le critère généralement utilisé lors de l'apprentissage en sera un de reconstruction parfaite.

### 3.2.1 Extraction de composantes

Pour l'utilisation du réseau dans le but d'extraire un vocabulaire de composantes utilisant un nombre prédéterminé de caractéristiques, on entraînera le réseau jusqu'à ce que l'erreur quadratique moyenne (calculée pour chaque stimulus entre $\mathbf{x}(0)$ et $\mathbf{x}(t)$) ne présente aucune amélioration d'un bloc d'apprentissage au suivant. On pourra alors trouver le dit vocabulaire en observant chacune des colonnes de la matrice de reconstruction.

### 3.2.2 Développement d'une mémoire d'exemplaires

Si le réseau vise une représentation parfaite des exemplaires en mémoire, accompagnée du développement d'un vocabulaire de composantes pouvant être directement exploré (matrice **V**), on présentera des exemplaires au réseau jusqu'à ce que la matrice **V** permette de les reconstruire parfaitement. Le rappel se fera de façon identique à une mémoire autoassociative récurrente : un stimulus d'entrée sera fourni au réseau, et itérera dans celui-ci à l'aide des matrice de poids finales, jusqu'à ce que la reconstruction récupérée soit exactement la même lors de deux itérations consécutives ($\mathbf{x}(t+1) = \mathbf{x}(t)$). Les rappels bruités et à base de stimuli incomplets suivront également la même procédure.

### 3.2.3 Catégorisation et identification autonomes

Dans le cas d'une utilisation pour la catégorisation et l'identification non-supervisée, l'apprentissage se fera jusqu'à ce que les représentations comprimées soient stables d'un bloc à l'autre (différence entre $\mathbf{y}(0)$ et $\mathbf{y}(t)$ pour tous les stimuli). Lors du rappel, on pourra explorer le paysage catégoriel, en effectuant un rappel itératif. Chaque stimulus itérera dans le réseau jusqu'à l'atteinte d'une compression stable ($\mathbf{y}(t+1) = \mathbf{y}(t)$). Les stimuli qui

convergeront vers la même représentation comprimée seront considérés comme indistincts pour le réseau, et par conséquent, étant membres d'une catégorie commune.

Un processus d'identification sera possible lorsque chaque stimulus convergera vers une représentation comprimée différente. Ainsi, chaque « catégorie » contiendra un seul membre, et cette absence de réels regroupements constitue la base d'un processus d'identification, où les objets en mémoire peuvent tous être distingués, sans nécessairement que les mémoires d'exemplaires soient perceptivement optimales.

## 3.3 Conclusion

Ce chapitre visait à présenter les spécifications techniques du modèle, ainsi que la logique sous-jacente à celui-ci. Plusieurs utilisations potentielles perceptivo-cognitives ont été proposées pour FEBAM. Il est à noter que pour la réalisation de toutes ces tâches, on utilisera qu'une seule architecture, une seule règle d'apprentissage, et une seule règle de transmission. Seuls le nombre d'unités et le critère d'arrêt du réseau seront variés.

Le prochain chapitre présente plusieurs implémentations possibles des différents mécanismes décrits ci-haut. On y trouvera des simulations de type « apprentissage machine », et des comparaisons de performance avec plusieurs algorithmes existants. Également, plusieurs réplications qualitatives de phénomènes perceptivo-cognitifs existants seront présentés.

# CHAPITRE IV

# FEBAM : EXPLORATION ET VALIDATION DES CARACTÉRISTIQUES TECHNIQUES ET COGNITIVES

Le présent chapitre vise à montrer la polyvalence du réseau FEBAM en ce qui a trait aux tâches d'inspiration cognitive et d'apprentissage machine. Trois classes de comportement seront explorées, soit l'extraction de caractéristiques, le développement de mémoires d'exemplaires parfaits et la catégorisation (et identification) non-supervisée.

La polyvalence du réseau est rendue possible par le double traitement des stimuli, qui est effectué simultanément. Tel que mentionné, l'extraction de caractéristiques et le processus menant à un rappel parfait des patrons d'entrée seront effectués à l'aide de la matrice de reconstruction. Le processus de catégorisation non-supervisée, et par ricochet, le processus d'identification non-supervisé, se feront à l'aide de la matrice de compression.

Il est important de mentionner que pour effectuer la totalité des tâches présentées dans ce chapitre, aucune modification ne sera faite au type d'architecture, ainsi qu'aux règles de transmission et d'apprentissage. Les qualités du réseau sont donc des qualités dites « de base » et ne nécessitent aucun ajout aux postulats déjà mentionnés.

## 4.1 Extraction de caractéristiques

Il a été précédemment déterminé (Chapitres 1/2) qu'un modèle perceptivo-cognitif devait être en mesure de créer son propre code réduit et son propre vocabulaire de composantes perceptuelles iconiques (Goldstone et al., 1998). Deux indices nous permettent de croire que FEBAM pourra effectuer cette tâche. Premièrement, l'utilisation d'une fonction d'erreur similaire à l'analyse en composantes principales non-linéaires permet de présumer que FEBAM possède des propriétés d'extraction de caractéristiques comparables à cette

classe de réseaux. Évidemment, ces capacités ne sont pas aussi contraintes que dans les réseaux connus, puisque l'absence de connexions latérales dans les couches du réseau FEBAM ne permet pas de garantir que les composantes extraites sont orthogonales ou même indépendantes. Comme il sera montré, cette absence de contraintes constitue en fait un avantage pour le réseau, améliorant la précision et la flexibilité de l'encodage. Deuxièmement, l'architecture de FEBAM est directement liée à celle d'un auto-encodeur. L'un des avantages d'interpréter FEBAM comme étant un auto-encodeur autonome est que l'identité visuelle des composantes extraites peut être directement récupérée en illustrant le contenu de la matrice de reconstruction[28]. Par opposition, la méthode générale utilisée pour illustrer des caractéristiques dans une RAM serait d'appliquer une analyse en vecteurs et valeurs propres à la matrice de poids finale (ceci fait du sens seulement lorsque la matrice de poids est obtenue de façon linéaire; BSB : Anderson et al., 1977; EIDOS : Bégin et Proulx, 1996).

### 4.1.1 Étude comparative : Compression et reconstruction d'images

Une façon de déterminer la validité et la qualité du processus de réduction dimensionnelle de FEBAM est d'effectuer une tâche classique de type « analyse en composantes principales », en l'occurrence la compression et reconstruction d'images en tons de gris (Haykin, 1994). Ici, les performances qualitatives et quantitatives de FEBAM seront comparées à celles d'algorithmes non-supervisés et généralement utilisés pour une telle tâche, soit des réseaux de neurones à base d'analyses en composantes principales (PCA) linéaire et non-linéaire, ainsi qu'un algorithme d'analyse en composantes indépendantes (ICA). Évidemment, le but avoué des simulations n'est pas de fournir un algorithme de compression plus performant, mais bien de proposer un modèle perceptivo-cognitif effectuant plusieurs types de tâches en utilisant les mêmes principes définitoires. Il est toujours possible de rendre un modèle ultra-performant en l'adaptant aux caractéristiques de la tâche. Cependant, tout gain en performance est généralement lié à une perte de généralité. La série de simulations effectuée dans ce chapitre sert plutôt à montrer la polyvalence et le potentiel du modèle en ce

[28] L'utilité des deux matrices séparées prend tout son sens ici. Dans les auto-encodeurs, la compression et la reconstruction se font à l'aide d'une seule matrice et de sa transposée. Pour cette classe de réseaux, l'identité des caractéristiques n'est généralement pas recherchée; seule la fonction de compression présente un intérêt.

qui a trait aux phénomènes perceptivo-cognitifs, et à valider le choix de l'architecture récurrente et des règles utilisées.

4.1.1.1 Méthodologie

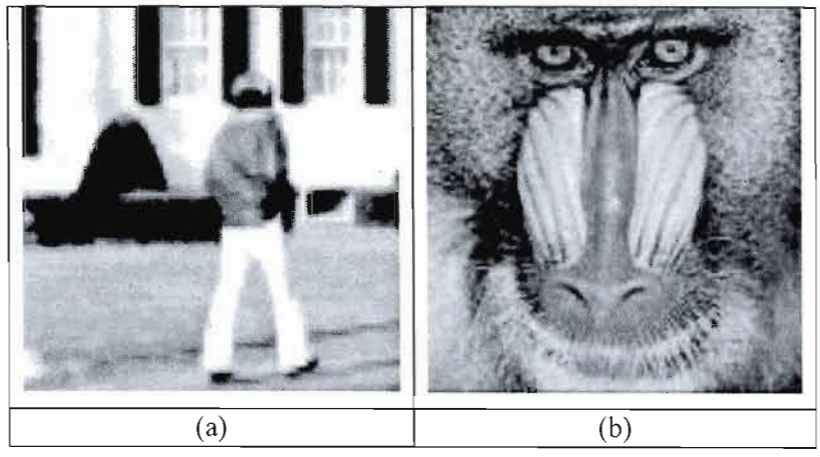


**Figure 4.1** Images en tons de gris utilisées pour (a) l'apprentissage; (b) la généralisation.

L'image utilisée lors de la phase d'apprentissage est illustrée à la Figure 4.1 (partie gauche). Cette image a une dimension de 128 pixels par 128 pixels, pour un total de 16384 pixels. Chaque image fut codée en une matrice de tons de gris (dont les valeurs entières vont de 0 à 255). Pour respecter le type de code requis par les algorithmes, ces valeurs ont ensuite été recodées (par transformation linéaire) pour suivre une échelle de valeurs continues contenue dans l'intervalle [-1, 1].

Puisque l'utilisation de vecteurs d'entrée d'une taille de 15376 positions provoquerait des calculs interminables, la matrice résultante fut divisée en vecteurs d'entrée de 25 dimensions, chaque vecteur représentant une fenêtre avec chevauchement de 5 pixels par 5 pixels[29]. Cinq unités de compression furent utilisées pour la simulation. Les paramètres de FEBAM furent fixés aux valeurs suivantes : le paramètre général de transmission fut fixé à $\delta = 0.1$ (donc entre 0 et 0.5, tel que requis), et le paramètre général d'apprentissage fut fixé à

[29] Ce choix, bien qu'arbitraire, constitue un compromis raisonnable entre une bonne intégration de l'information topologique voisine et un bon degré de clarté de l'image. Le choix n'a cependant pas vraiment d'impact au point de vue des comparaisons entre modèles, puisque tous sont testés à l'aide des mêmes patrons d'entrée.

$\eta$ = 0.005, en accord avec la condition énoncée à l'équation 3.5 (valeur maximale: 0.025). Pour obtenir une estimation robuste de la performance du réseau, 25 simulations séparées furent réalisées à l'aide du même ensemble de vecteurs d'entrée. Pour chaque simulation, la procédure d'apprentissage fut la suivante :

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]);
1. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de vecteurs;
   b. Réalisation d'un cycle dans le réseau, tel qu'illustré à la Figure 3.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 3.3 et 3.4;
2. Répétition de l'étape 1 jusqu'à l'atteinte d'un seuil d'erreur quadratique inférieur à 0.005 (la répétition d'un même vecteur d'entrée est permise). L'erreur quadratique (EQ) est calculée d'après la qualité de la reconstruction du stimulus d'entrée effectuée par le réseau (en utilisant les caractéristiques extraites). Ainsi le calcul exact est :

$$EQ = \left[ (\mathbf{x}(1) - \mathbf{x}(0))^T (\mathbf{x}(1) - \mathbf{x}(0)) \right] / M \qquad 4.1$$

où $M$ représente le nombre d'unités de la couche d'entrée.

La procédure de rappel fut effectuée comme suit :

1. Essai de rappel :
   a. Sélection de l'un des vecteurs d'entrée;
   b. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $\mathbf{x}(t+1) = \mathbf{x}(t)$);
   c. Sélection du pixel central dans la fenêtre de 5 par 5 pixels résultante (vecteur de 25 positions divisé en 5);
2. Répétition de l'étape 1 pour chaque patron (vecteur) de l'ensemble;
3. Reconstruction de l'image à l'aide des pixels sélectionnés.

#### 4.1.1.2 Résultats : Phase d'apprentissage[30]

Après 40000 essais d'apprentissage, le réseau atteint le seuil d'erreur quadratique minimal. Ainsi, la matrice de reconstruction constitue une solution quasi-optimale à l'équation 3.4. La matrice résultante de l'une des simulations est illustrée à la Figure 4.2. On

[30] Les résultats présentés dans cette section sont à titre illustratif, et ne concernent qu'une seule simulation. Toutes les simulations ont donné des résultats similaires, en termes du nombre d'essais d'apprentissage et du type de composantes développées.

peut observer que le réseau développe naturellement des composantes s'apparentant à des traits horizontaux, verticaux et diagonaux. Ces traits sont tout de même quelque peu flous, attestant de la flexibilité nécessaire pour la reconstruction adéquate de l'image en tons de gris.

**Figure 4.2** Graphes de densité à contours arrondis représentant les poids de connexion finaux pour la matrice **V**. Chaque image est équivalente à une colonne de la matrice (partitionnée en 5 lignes de 5 pixels) et représente une composante utilisée par le réseau pour reconstruire les images.

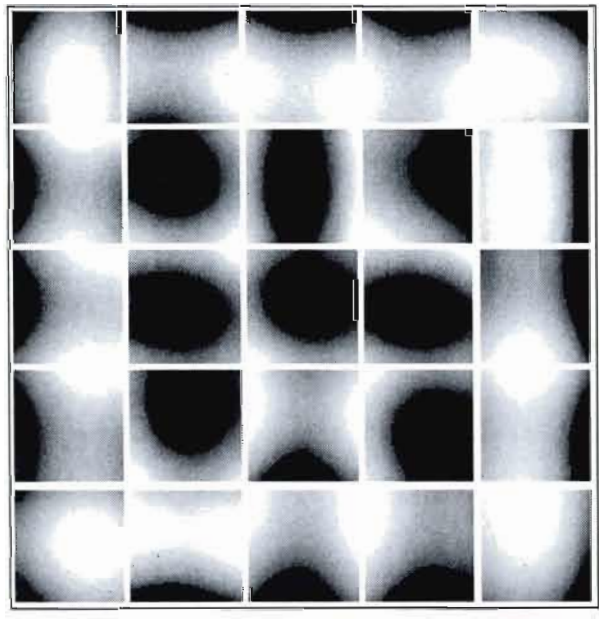

**Figure 4.3** Détecteurs développés lors de l'apprentissage. Ces détecteurs indiquent à quelle information spatiale chaque unité de la couche d'entrée est devenue sensible suite à l'apprentissage.

Il est possible également de déterminer à quelle information spatiale les unités de la couche d'entrée sont devenus sensibles suite à l'apprentissage. La Figure 4.3 illustre les 25

détecteurs développés à la couche d'entrée. Ces détecteurs ont été obtenus en multipliant ensemble des éléments des matrices **V** et **W**, suivant la formule suivante :

$$\forall j, j = i, \mathbf{V}_j \times (\mathbf{W}^T)_i \qquad 4.2$$

où $j$ représente une colonne de la matrice **V**, et $i$ représente une ligne de la matrice **W** transposée. Ici aussi, on peut déceler une sensibilité aux traits de différentes orientations (horizontale, verticale, diagonale) mais aussi une sensibilité à la position spatiale des pixels à l'intérieur de la fenêtre réduite.

#### 4.1.1.3 Résultats : Reconstruction d'images (rappel)

La performance de FEBAM fut comparée à celle de trois autres modèles autonomes, choisis comme représentants adéquats de leur classe respective : APEX (Diamantaras et Kung, 1996), un réseau de neurones à base de PCA linéaire; nPCA, (Karhunen, Pajunen et Oja, 1998), un réseau de neurones à base de PCA non-linéaire; et fastICA (Hyvarinen et Oja, 2000), un algorithme d'analyse en composantes indépendantes (ICA). La Figure 4.4 permet de comparer les performances de reconstruction des différents algorithmes utilisés. Visuellement, il est facile de déterminer que FEBAM et fastICA sont les algorithmes permettant de mieux reproduire les subtilités de l'image originale lors de la reconstruction. Les réseaux à base d'analyse en composantes principales ont tendance à accentuer l'utilisation de valeurs extrêmes (noir et blanc) au détriment des tons de gris de l'image originale. De façon théorique, ce résultat est sensé, puisque le PCA (du moins lorsqu'effectuée par APEX) vise théoriquement à éliminer les corrélations entre les composantes, et donc à les rendre visuellement moins semblables, en utilisant un processus dynamique menant à une orthogonalisation de Gram-Schmidt.

La mesure quantitative de comparaison utilisée est le ratio entre le signal maximal et le bruit (*peak signal-to-noise ratio*, ou *PSNR*; Garnett, Huegerich, Chui et He, 2005), calculé comme suit :

| *Modèle* | *Reconstruction (une simulation)* | *PSNR* | *Généralisation (une simulation)* | *PSNR* |
|---|---|---|---|---|
| (a) FEBAM |  | 31.98 |  | 25.93 |
| (b) APEX |  | 17.74 |  | 22.19 |
| (c) nPCA |  | 21.39 |  | 20.37 |
| (d) fastICA |  | 31.13 |  | 26.34 |

**Figure 4.4** Résultats de reconstruction et de généralisation pour les différents algorithmes utilisés : (a) FEBAM; (b) APEX, un réseau de neurones à base de PCA linéaire (Diamantaras et Kung, 1996); (c) nPCA, un réseau de neurones à base de PCA non-linéaire (Karhunen, Pajunen et Oja, 1998); (d) fastICA, un algorithme d'analyse en composantes indépendantes (Hyvarinen et Oja, 2000). Les valeurs de *PSNR* rapportées sont des moyennes basées sur 25 simulations.

$$PSNR(z) = 10 \log_{10} \left( \frac{\sum_{i,j=1}^{ligne,colonne} (255)^2}{\sum_{i,j=1}^{ligne,colonne} (z_{i,j} - o_{i,j})^2} \right) \quad 4.3$$

où $z$ est l'image reconstruite et $o$ est l'image originale. Un PSNR élevé dénote une qualité de reconstruction accrue.

En termes quantitatifs, en ce qui a trait à la reconstruction de l'image originale (Figure 4.4), la performance moyenne de FEBAM est supérieure à celle d'APEX et de nPCA, mais équivalente à celle de fastICA. La capacité de généralisation à partir des composantes extraites fut également évaluée. La procédure de rappel fut répétée en utilisant une seconde image avec des propriétés statistiques distinctes (Figure 4.1, partie droite). Cette image fut également divisée en vecteurs d'entrée de 25 dimensions, chaque vecteur représentant une fenêtre avec chevauchement de 5 pixels par 5 pixels. Les résultats (Figure 4.4, partie droite), en termes de rang comparatifs, sont les mêmes. FEBAM et fastICA présentent des résultats très similaires, et supérieurs aux autres réseaux testés. Il semble donc que les capacités d'extraction de caractéristiques de FEBAM soient valides au niveau de l'apprentissage machine.

## 4.2 Développement non-supervisé d'une mémoire d'exemplaires

Une mémoire associative de type récurrent, même si on lui ajoute une couche de compression, devrait être en mesure de continuer à présenter certaines caractéristiques cognitives, telle la possibilité de parfaitement reconstruire les stimuli. Cette capacité correspond, en psychologie cognitive, à un processus « extrême » de mémoire, qui mènerait au développement d'une mémoire d'exemplaires parfaite. En effet, pour un système cognitif humain, le critère initial de réussite d'une mémoire d'exemplaires devrait être la capacité de discriminer les objets entre eux (retour au principe de différentiation de Gibson), mais pas nécessairement de pouvoir permettre aux humains de reproduire dans les moindres détails l'image d'entrée à partir de la mémoire.

FEBAM présente l'avantage cognitif principal de pouvoir comprimer la taille des représentations dans le réseau. Une question pertinente est de savoir si cet avantage se fait au détriment d'autres caractéristiques techniques héritées de NDRAM (Chartier et Proulx, 2005) et de plusieurs autres modèles du type RAM. Ces propriétés incluent la tolérance au rappel bruité ou incomplet, et la résistance aux attracteurs nuisibles (*spurious attractors*). Les prochaines simulations visent donc principalement à comparer les performances de FEBAM avec celles de la mémoire autoassociative récurrente dont ce modèle est dérivé, en l'occurrence NDRAM. Elles visent également à explorer les comportements du réseau en ce qui a trait à la mémorisation parfaite d'exemplaires par extraction de caractéristiques, sous différentes conditions d'apprentissage et de rappel.

### 4.2.1 Étude de comportement : Apprentissage et rappel de stimuli bipolaires

Cette première simulation vise à étudier les comportements du modèle lors d'un apprentissage et d'un rappel effectué dans des conditions normales. La possibilité de réduire la taille des représentations dans le réseau devrait mener à une certaine économie cognitive lorsque l'on compare avec une RAM, qui, par convention, utilise un nombre d'unités égal au nombre de pixels.

Également, une attention particulière sera portée à l'effet des conditions de départ. En effet, puisqu'avec FEBAM, les matrices de poids sont initialisées aléatoirement, il est possible que la réussite de la tâche soit étroitement liée à un ensemble spécifique de poids, ce qui serait catastrophique pour le réseau.

#### 4.2.1.1 Méthodologie

Les stimuli utilisés pour cette étude sont illustrés à la Figure 4.5. Il s'agit de versions matricielles (5 pixels par 7 pixels) des lettres majuscules A, B, C, D, E, H, I, N, O, R, S et T. Les pixels noirs et blancs représentent des valeurs respectives de 1 et -1. Pour l'entrée au réseau, chaque stimulus fut transformé en un vecteur de 35 positions. Aucun prétraitement ne fut effectué préalablement aux simulations. Ces stimuli, originalement utilisés par Chartier et Proulx (2005), ont été sélectionnés pour la variété des corrélations inter-stimuli, qui (en

valeur absolue) varient entre $r = 0.01$ (quasi-inexistante; Cohen, 1988) entre les lettres C et T, et $r = 0.84$ (très forte; Cohen, 1988) entre les lettres C et O.

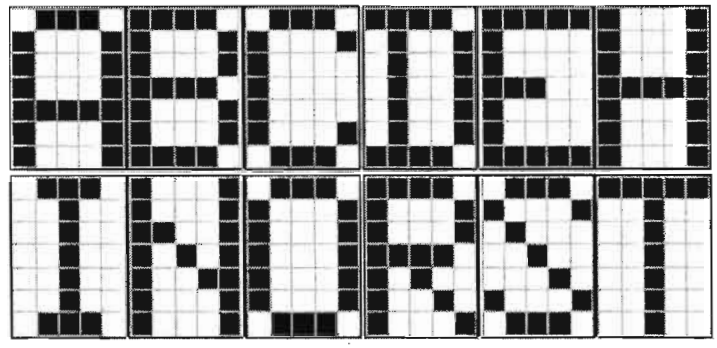

**Figure 4.5** Stimuli utilisés pour la simulation 4.2.1.

L'effet du nombre d'unités de compression dans le réseau (couche **y**) fut étudié. Ainsi, l'apprentissage fut répété en utilisant un nombre d'unités de compression variant entre 1 et 35. Pour chaque nombre d'unités, 200 simulations (utilisant des matrices de poids aléatoires initiales différentes) furent réalisées (pour un total de 7000 simulations).

Les paramètres de FEBAM furent fixés aux valeurs suivantes : le paramètre général de transmission fut fixé à $\delta = 0.1$, et le paramètre général d'apprentissage fut fixé à $\eta = 0.01$ (valeur maximale: 0.018). La procédure d'apprentissage se déroula comme suit:

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]);
1. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le réseau, tel qu'illustré à la Figure 3.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 3.3 et 3.4;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Calcul de l'erreur quadratique moyenne pour l'ensemble de stimuli;
4. Répétition des étapes 1 à 3 jusqu'à l'atteinte d'un seuil d'erreur quadratique moyen inférieur à $1 \times 10^{-6}$.

La procédure de rappel se déroula comme suit :

1. Essai de rappel :
   a. Sélection de l'un des 12 stimuli d'entrée;
   b. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $\mathbf{x}(t+1) = \mathbf{x}(t)$);
2. Répétition de l'étape 1 pour chaque stimulus de l'ensemble.

Lors du rappel, une reconstruction finale (stable) fut considérée comme acceptable si l'erreur quadratique entre le stimulus d'entrée $\mathbf{x}(0)$ et la reconstruction $\mathbf{x}(t)$ était inférieure à $1x10^{-6}$ (cette différence est invisible à l'œil nu). Un rappel fut déclaré « parfait » lorsque la reconstruction des 12 stimuli originaux respectait cette règle.

#### 4.2.1.2 Résultats

La Figure 4.6 montre que FEBAM peut parfaitement reconstruire l'ensemble de stimuli, à condition que le nombre d'unités de compression soit suffisant. Ici, 11 unités sont requises pour mener à un rappel parfait en tout temps[31]. Ainsi, il est clair que l'initialisation aléatoire des matrices de poids de connexion ne constitue pas un problème au point de vue de la stabilité du processus, en autant que le nombre d'unités de compression soit suffisant pour accomplir le travail[32]. Le processus semble en être un de type « tout ou rien ». À 6 unités, aucun rappel n'est parfait, et avec seulement 5 unités de plus, tous les rappels le sont.

La réduction dimensionnelle possible constitue donc une nette économie, tant au niveau de la taille des représentations, qu'à celui du nombre de connexions requises. En effet, pour effectuer le même travail, NDRAM (ou toute autre RAM) utilise 35 unités de représentation, et le nombre de connexions du réseau est égal à 1225 (35 x 35). FEBAM peut effectuer la tâche en mémorisant des représentations réduites sur 11 unités (économie représentationnelle : 68.6%), et le nombre de connexions nécessaires est alors égal à 770 (11 x 35 x 2 matrices; économie : 37.1%).

---

[31] On pourrait croire que le nombre théorique minimal d'unités de compression devrait être égal au nombre de stimuli. Ce serait le cas dans un modèle compétitif de type « winner-take-all ». Ici, le fait que le réseau puisse créer des états positionnés autrement qu'aux vertex de l'hyperespace fait que le nombre d'unités minimal est dépendant des corrélations interstimuli.

[32] D'autres simulations non rapportées ici ont été réalisées en initialisant les poids avec des valeurs de l'intervalle [-1, 1]. Le fait d'utiliser un intervalle plus large n'a eu aucune influence sur le nombre de rappels parfaits. Cela n'a fait que ralentir l'atteinte du critère de réussite à l'apprentissage.

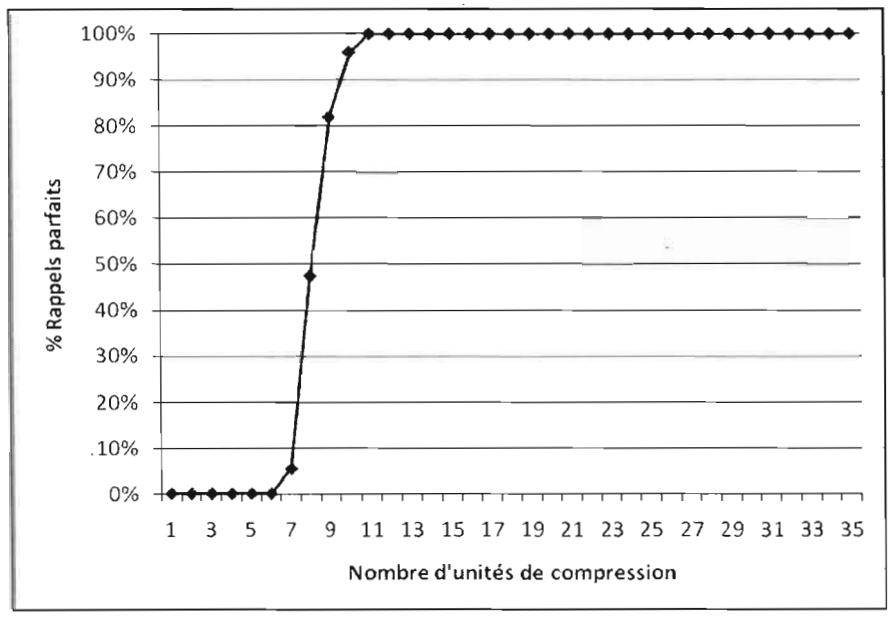


**Figure 4.6** Pourcentage de rappels parfaits en fonction du nombre d'unités de compression (200 simulations par nombre d'unités)

Cette économie représentationnelle se fait cependant au profit d'un ralentissement de l'apprentissage. En effet, tel que le montre la Figure 4.7, le nombre de blocs d'apprentissage nécessaire augmente lorsque l'on comprime les représentations de façon plus marquée. Lorsqu'aucune compression n'est effectuée, la rapidité de l'apprentissage est comparable à celle de NDRAM (FEBAM : 38.68 blocs en moyenne; NDRAM : 46.01 blocs[33]).

Il est à noter toutefois que la relation entre le nombre d'unités de compression et le nombre de blocs d'apprentissage requis suit fidèlement une courbe de puissance[34]. Ceci est un point important, puisqu'il indique que le ralentissement n'est marqué que lorsque la compression est quasi-maximale. Par exemple, lorsque l'on passe de 35 à 17 unités, le temps d'apprentissage double, mais il double encore lorsque l'on passe de 17 à 11 unités (une différence de seulement 6 unités). Il serait donc tout de même possible de réaliser une « compression neuronale » sans considérablement ralentir l'apprentissage. À titre d'exemple,

[33] Cette moyenne a été obtenue en simulant le réseau NDRAM 200 fois, en utilisant une présentation par blocs. Les paramètres utilisés sont les mêmes que pour FEBAM, soit $\delta = 0.1$ et $\eta = 0.001$.

[34] Meilleure équation estimée : $y = 207.68\ x^{-0.511}$, où x représente la taille de la couche de compression, et y, le nombre de blocs requis; coefficient de détermination : $r^2 = 0.9883$. Cette équation exacte ne vise pas une quelconque interprétation cognitive, mais permet d'indiquer que l'apprentissage ne ralentit qu'en cas de compression maximale.

on peut évaluer qu'une compression en 25 unités (économie représentationnelle : 28.5%), ne nécessiterait que 13.01 blocs d'apprentissage supplémentaires.

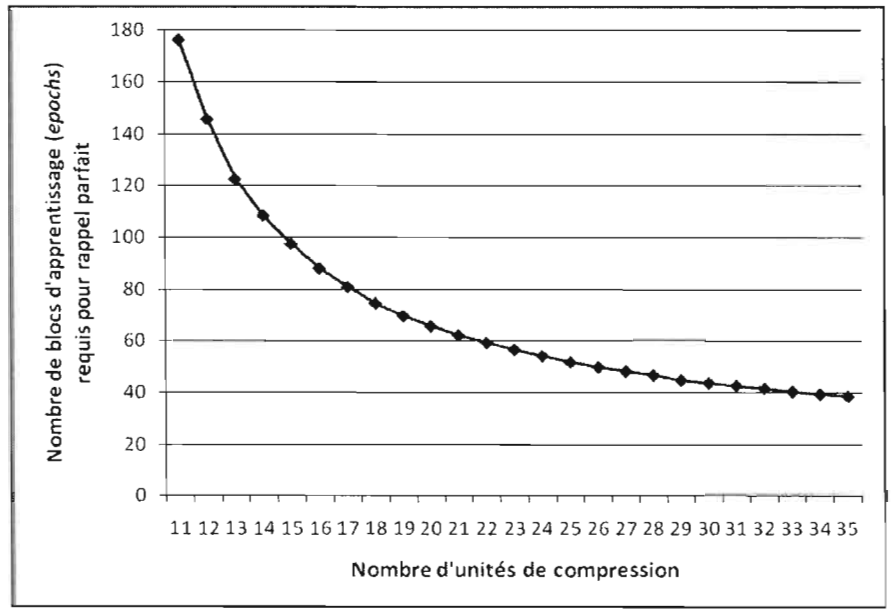


**Figure 4.7** Nombre de blocs d'apprentissage (epochs) nécessaires pour un rappel parfait, en fonction du nombre d'unités de compression (200 simulations par nombre d'unités). L'erreur-type suit également une courbe de puissance - erreur-type maximale (11 unités): 2.35 blocs, erreur-type minimale (35 unités) : 0.17 blocs.

Du point de vue des représentations comprimées, il serait possible qu'avec un niveau de compression maximal, le réseau utilise des représentations localistes, où chaque stimulus serait en fait associé à une seule composante de la matrice **V** (ou une unité de la couche **y**). L'étude des composantes et des représentations au rappel pourrait nous indiquer si c'est le cas, ou si le réseau utilise des représentations comprimées distribuées, où la contribution de chacune des composantes varie d'un stimulus à l'autre. La Figure 4.8 présente un exemple de simulation complète réalisée à l'aide de 12 unités de compression.

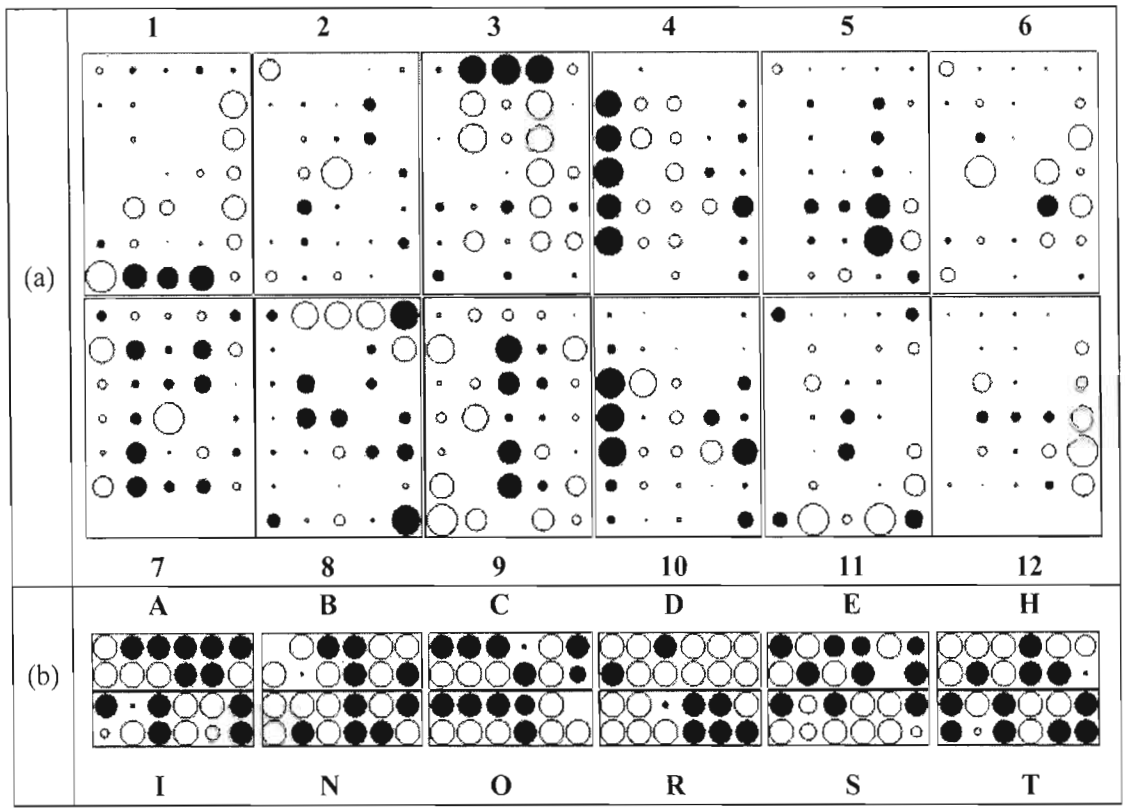


**Figure 4.8** (a) Diagrammes à bulles (Cousineau, Lacroix, et Hélie, 2003) représentant les composantes développées par le réseau. Chaque diagramme est équivalent à une colonne de la matrice **V** (partitionnée en 7). Les cercles complets représentent des poids de connexion positifs, et les cercles vides les poids négatifs. La taille de chaque cercle est proportionnelle à la valeur du poids; les plus grands cercles représentent des valeurs égales à -1 ou 1. (b) Diagrammes à bulles pour les représentations comprimées récupérées lors du rappel. Chaque cercle représente la force d'activation de chacune des 12 composantes pour la reconstruction de chaque lettre. Un cercle noir signifie qu'une composante est activée, alors qu'un cercle blanc signifie qu'elle est inhibée.

La Figure 4.8(a) indique clairement que le réseau utilise des représentations distribuées. Au point de vue des 12 composantes développées par le réseau, aucune des composantes illustrées ne représente l'un des stimuli originaux. On peut plutôt y déceler un ensemble de composantes dont certaines sont directement interprétables, tels les traits horizontaux (composantes 1 et 3) ou verticaux (composante 9) composant les lettres. Fait intéressant, l'utilisation du codage [-1,1] permet également le développement de composantes dénotant la présence (cercles noirs de grande taille) et l'absence (cercles vides de grande taille) de pixels simultanément. La Figure 4.8(b) constitue une autre preuve de distribution des

représentations : en effet, il est ici montré que les représentations comprimées de chacune des lettres apprises utilisent simultanément plus d'une composante développée, et que certaines lettres utilisent des composantes communes. Par exemple, on peut déceler que la lettre I est principalement reconstruite grâce aux composantes 1 (trait horizontal), 3 (trait horizontal) et 9 (trait vertical).

Il semble donc qu'à la base, FEBAM rencontre plusieurs objectifs désirés, tels que l'économie en termes de stockage, le maintien de représentations distribuées, et la capacité de performer au même niveau qu'une mémoire autoassociative récurrente lorsque le nombre d'unités est égal dans les deux couches. Il est également rassurant de voir que les conditions de départ des matrices de poids ne constituent nullement un facteur déterminant pour la performance du réseau.

### 4.2.2 Étude de comportement : Apprentissage et rappel de stimuli en tons de gris

Cette étude constitue une réplication de la précédente, mais utilisant un ensemble de stimuli distincts. L'intérêt d'utiliser un second ensemble de stimuli est double. Premièrement, il permettra d'illustrer la capacité de FEBAM de traiter des stimuli à valeurs continues, ce qui en fait est lié à la règle d'apprentissage, et constitue un héritage des modèles NDRAM et BHM. Deuxièmement, les corrélations inter-stimuli réduites devraient permettre une plus grande économie d'espace en mémoire, puisque le nombre d'unités de compression nécessaire pour l'encodage sera réduit. En effet, le réseau aura moins de composantes à développer pour séparer les stimuli dans l'espace multidimensionnel.

#### 4.2.2.1 Méthodologie

Les stimuli utilisés pour cette étude sont illustrés à la Figure 4.9. Il s'agit de versions matricielles (16 pixels par 16 pixels) d'icônes informatiques en tons de gris. Les pixels sont représentés par des valeurs continues incluses dans l'intervalle [-1, 1], où les pixels noirs et blancs ont des valeurs respectives de 1 et -1. Pour l'entrée au réseau, chaque stimulus fut transformé en un vecteur de 256 positions. Aucun prétraitement ne fut effectué préalablement

aux simulations. Les corrélations inter-stimuli (en valeur absolue) variaient entre $r = 0.01$ (quasi-inexistante; Cohen, 1988) et $r = 0.34$ (moyenne; Cohen, 1988).

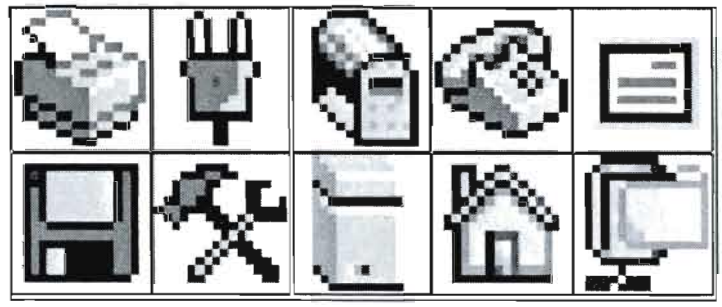

**Figure 4.9** Stimuli utilisés pour la simulation 4.2.2.

Le nombre d'unités de compression dans le réseau (couche **y**), fut varié, allant de 1 à 32 (intervalles de 1) et de 32 à 256 (intervalles de 16). Pour chaque nombre d'unités, 200 simulations, utilisant des matrices de poids aléatoires initiales différentes, furent réalisées, pour un total de 9200 simulations. Le paramètre général de transmission fut fixé à $\delta = 0.1$, et le paramètre général d'apprentissage fut fixé à $\eta = 0.001$ (valeur maximale: 0.002). Les procédures d'apprentissage et de rappel, ainsi que les critères de reconstruction acceptable, furent identiques à ceux de l'étude précédente.

### 4.2.2.2 Résultats

La Figure 4.10 montre que dans le cas de ce second ensemble, FEBAM peut parfaitement reconstruire l'ensemble de stimuli à l'aide d'un minimum de 21 unités de compression. Une fois de plus, l'initialisation aléatoire des matrices de poids de connexion n'est nullement problématique; lorsqu'un nombre d'unités de connexions est suffisant pour le rappel parfait, tout nombre d'unités supérieur mène également à des rappels parfaits. Point de vue économie, pour effectuer le même travail, NDRAM (ou toute autre RAM) utilise 256 unités de représentation, et le nombre de connexions du réseau est égal à 65536. FEBAM peut effectuer la tâche en mémorisant des représentations réduites sur 21 unités (économie représentationnelle : 91.8%), et le nombre de connexions nécessaires est alors égal à 10752 (21 x 256 x 2 matrices; économie : 83.6%).

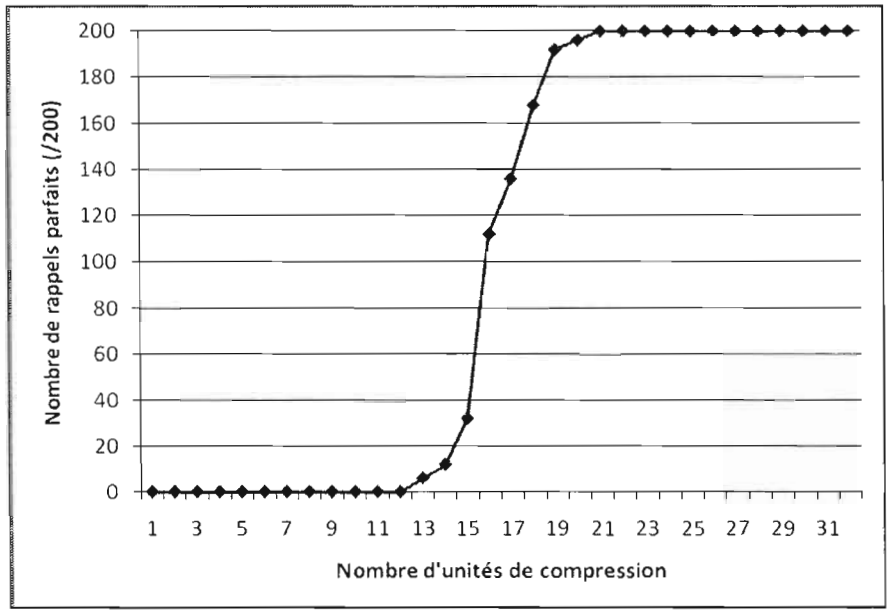


**Figure 4.10** Pourcentage de rappels parfaits en fonction du nombre d'unités de compression (200 simulations par nombre d'unités).Lorsque le réseau utilise plus de 32 unités, le pourcentage de rappels parfaits est toujours égal à 100%.

La Figure 4.11 présente la relation entre le nombre de blocs d'apprentissage nécessaire pour assurer un rappel parfait et le nombre d'unités de compression. Il semble acquis qu'en général, la relation entre ces deux variables suit réellement une courbe de puissance[35]. Si l'on accepte de doubler le nombre de blocs d'apprentissage, l'économie représentationnelle est alors de 62.5% (96 unités de compression). Ainsi, lorsque les stimuli sont plus faciles à distinguer, FEBAM réussit donc à profiter de la baisse des corrélations, et permet une compression encore plus économique cognitivement.

[35] Meilleure équation estimée : $y = 225.72\ x^{-0.662}$, où x représente la taille de la couche de compression, et y, le nombre de blocs requis; coefficient de détermination : $r^2 = 0.995$.

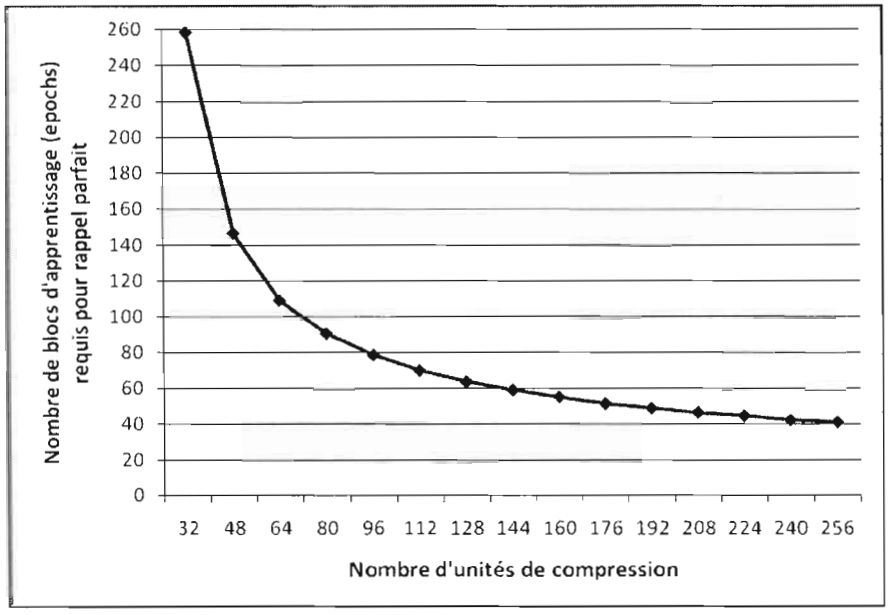


**Figure 4.11** Nombre de blocs d'apprentissage (epochs) nécessaires pour un rappel parfait, en fonction du nombre d'unités de compression (200 simulations par nombre d'unités). L'erreur-type suit également une courbe de puissance - erreur-type maximale (32 unités) : 3.19 blocs, erreur-type minimale (256 unités) : 0.25 blocs. Lorsque le nombre d'unités est maximal, le nombre de blocs requis est similaire à celui pour NDRAM (36.74 blocs)[36].

### 4.2.3 Étude comparative : Rappel bruité (stimuli bipolaires)

En général, dans une RAM, lorsqu'un stimulus est présenté au réseau lors du rappel, ce dernier itère dans l'espace stimuli jusqu'à la stabilisation sur un attracteur donné. Le processus itératif permet au vecteur d'entrée de progressivement modifier sa direction et sa norme, jusqu'à ce que la pointe du vecteur s'arrête sur un attracteur (Hopfield, 1982). Ce processus est rendu possible par la boucle de récurrence. Ceci est d'autant plus important lorsque le vecteur d'entrée a subi une certaine dégradation; la boucle permet ainsi de filtrer le bruit pour ainsi retrouver l'un des stimuli de l'ensemble original.

Chartier et Proulx (2005) ont montré que NDRAM présentait des performances de rappel bruité supérieur à d'autres modèles de la même classe (Bégin et Proulx, 1996; Diederich et Opper, 1987; Kanter et Sompolinsky, 1987; Storkey et Valabregue, 1999), faisant de ce réseau un excellent étalon de comparaison en ce qui a trait à la résistance des

[36] Cette moyenne a été obtenue en simulant le réseau NDRAM 200 fois, en utilisant une présentation par blocs. Les paramètres utilisés sont les mêmes que pour FEBAM, soit $\delta = 0.1$ et $\eta = 0.0001$.

RAM au rappel bruité. On comparera donc FEBAM à NDRAM pour cette raison, et aussi parce que les deux réseaux sont mathématiquement liés.

Dans FEBAM, la présence de deux espaces séparés et asymétriques (un espace par matrice) n'empêche pas le principe de récurrence, déjà utilisé pour le rappel jusqu'ici; la performance de filtration du bruit pourrait cependant être inférieure, puisque le réseau ne vise pas la stabilité, mais bien la qualité de reconstruction. Le but principal de la simulation suivante était d'explorer les effets combinés du bruit au rappel et de la compression des représentations. Les performances quantitatives du modèle seront une fois de plus comparées avec celles de NDRAM.

#### 4.2.3.1 Méthodologie

Les stimuli, les paramètres et la procédure utilisés pour l'apprentissage sont identiques à ceux de la simulation 4.2.1. La différence entre la simulation 4.2.1 et celle-ci se situe plutôt au niveau du rappel. En effet, un vecteur suivant une distribution normale avec moyenne 0 et écart-type $\sigma$, fut ajouté à chacun des stimuli originaux avant d'effectuer le rappel (Figure 4.12). Suivant Bégin et Proulx (1996), lorsque les vecteurs d'entrée n'ont pas à être normalisés pour le traitement, $\sigma$ représente le rapport entre la norme du vecteur aléatoire et celle du vecteur d'origine. On peut donc directement considérer $\sigma$ comme la proportion de bruit ajouté au stimulus[37].

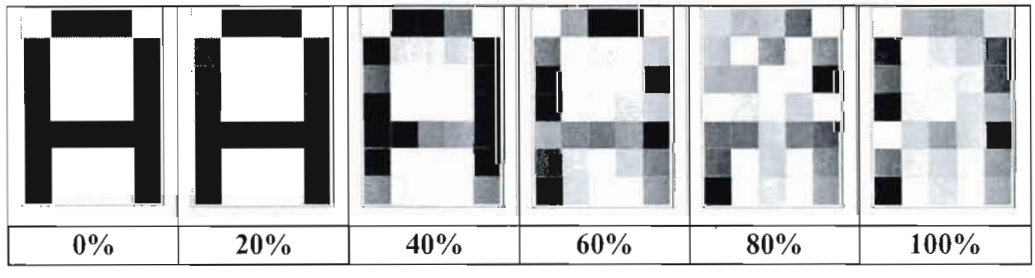


**Figure 4.12** Graphes de densité illustrant l'ajout d'un vecteur de bruit à la lettre A. Les nombres représentent le pourcentage de bruit ajouté.

[37] Contrairement à la procédure décrite dans Bégin et Proulx (1996) pour EIDOS, les vecteurs d'entrée n'ont pas à être normalisés avec FEBAM et NDRAM.

La procédure de rappel fut donc légèrement modifiée, et se détaille comme suit :

1. Essai de rappel :
   a. Sélection de l'un des 12 stimuli d'entrée;
   b. Addition d'un vecteur de bruit aléatoire **b** au stimulus original, $\mathbf{b} \sim N(0, \sigma)$;
   c. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $x(t+1) = x(t)$);
   d. Répétition des étapes *b* et *c* pour 240 rappel bruités;
   e. Calcul de la proportion de rappel parfait pour le stimulus;
2. Répétition de l'étape 1 pour chaque stimulus de l'ensemble.

Le nombre d'unités de compression (entre 11 et 31, par intervalles de 4) [38], ainsi que le niveau de bruit au rappel (20%, 40%, 60%, 80%, 100%) furent variés. Pour chaque combinaison de niveau de bruit et de nombre d'unités, 100 simulations, utilisant des matrices de poids aléatoires initiales différentes, furent réalisées (pour un total de 3000 simulations), dans le but d'obtenir une estimation moyenne robuste du pourcentage de rappels parfait. Dans le but de limiter le nombre de simulations à réaliser pour tenir compte des multiples niveaux de chaque variable, le nombre d'unités de compression de FEBAM a été augmenté jusqu'à ce que la performance du réseau soit égale ou supérieure à celle de NDRAM pour tous les niveaux d'incomplétude (ce critère n'a jamais été atteint). Les vecteurs de bruit furent générés aléatoirement pour chaque simulation. Lors du rappel, une reconstruction finale (stable) fut considérée comme acceptable si l'erreur quadratique entre le stimulus d'entrée et la reconstruction était inférieure à $1x10^{-6}$.

[38] La simulation de FEBAM avec 35 unités présente peu d'intérêt, puisque la proposition du modèle vise à pouvoir effectuer la réduction dimensionnelle et la mémoire d'exemplaires de façon simultanée. C'est pourquoi elle n'a pas été réalisée.

4.2.3.2 Résultats

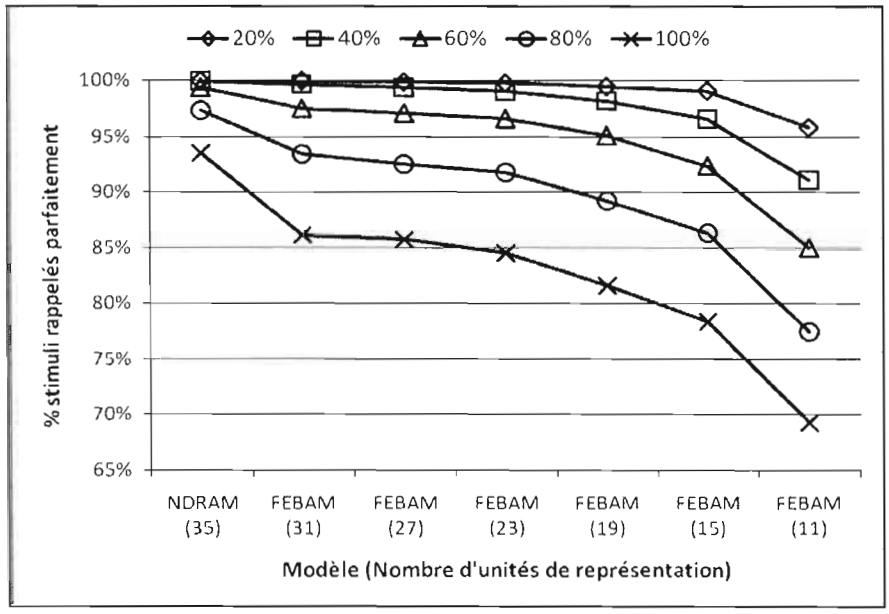


**Figure 4.13** Pourcentage de lettres rappelées parfaitement en fonction du nombre d'unités de compression et du niveau de bruit ajouté. Les résultats obtenus avec NDRAM (100 simulations utilisant les mêmes paramètres qu'avec FEBAM) sont également présentés pour comparaison. Chaque ligne représente un niveau de bruit différent[39].

La tolérance de FEBAM au bruit est une fonction directe du niveau de compression. En effet, pour obtenir un taux de rappel supérieur à 95%, on doit insérer un minimum de 11 unités lorsque le niveau de bruit est de 20%, 15 unités lorsqu'il est de 40%, et 19 unités pour 60%. Lorsque le niveau de bruit est supérieur à 60%, la dégradation de la performance est nette pour FEBAM, et l'impact du nombre d'unités de compression est clair et montre une décroissance monotone. À ce niveau de bruit, la réduction de performance est encore plus marquée lorsque l'on compare avec NDRAM, qui peut facilement encaisser jusqu'à 80% de bruit au rappel. Les performances à cette tâche pour FEBAM, même lors de l'utilisation d'un nombre minimal d'unités de compression, sont égales ou supérieures à celles des modèles comparés à NDRAM dans Chartier et Proulx (2005). À titre d'exemple, lors que l'on utilise

[39] Les barres d'erreur n'apparaissent pas sur les graphes pour les études 4.2.3, 4.2.4 et 4.2.5. L'écart-type n'a malheureusement pas été enregistré lors de la majorité des simulations du Chapitre 4. Dû au nombre de simulations réalisées dans chacune des conditions, et suite aux résultats d'autres études présentées dans cette thèse, on peut cependant croire que les barres d'erreur seraient pratiquement invisibles sur les graphes.

100% de bruit, FEBAM (avec 11 unités) montre une performance de rappel de 68.9%. En comparaison, les modèles linéaires optimaux (Diederich et Opper, 1987; Hanter et Sompolinsky, 1987; Storkey et Valabregue, 1999) montrent une performance avoisinant les 50%. EIDOS (Bégin et Proulx, 1996) performe mieux que FEBAM à cette tâche, mais seulement lorsque 23 unités de compression et moins sont utilisées.

Malgré la comparaison quelque peu défavorable avec NDRAM, cette étude montre une propriété intéressante de FEBAM. Une particularité de ce dernier, en ce qui a trait au rappel bruité, est liée à la sortie finale du réseau. En effet, FEBAM se stabilise rarement sur un vecteur non inclus dans l'ensemble original de stimuli, ce que NDRAM fait régulièrement. Plutôt que s'arrêter sur un vecteur non-désiré, FEBAM semble plutôt effectuer des erreurs de confusion, tout comme les humains. Dans le cas de l'ensemble à l'étude, les stimuli C et O sont responsables de la quasi- totalité des erreurs de confusion. Ceci est normal, compte tenu de la corrélation élevée entre ces deux vecteurs. Ainsi, en cas d'erreur, la grande majorité des reconstructions stables représentent parfaitement l'un ou l'autre des stimuli de l'ensemble original, mais pas nécessairement le bon stimulus. Ceci porte à croire que les espaces vectoriels créés par FEBAM contiennent moins d'attracteurs nuisibles (*spurious states*); ce phénomène sera étudié plus spécifiquement dans l'une des prochaines simulations.

Il faut aussi tenir compte du fait que l'étude de la résistance au bruit par ajout d'un vecteur, si elle constitue un standard de comparaison en apprentissage machine, fait peu de sens lorsque l'on s'intéresse aux processus cognitifs humains. En effet, malgré que l'on puisse postuler un certain niveau de bruit interne dans le système cognitif, ce bruit serait présent non seulement au rappel, mais aussi à l'apprentissage. Rainer et Miller (2000) ont d'ailleurs montré que l'apprentissage bruité augmentait la résistance au rappel lorsque les stimuli y sont dégradés. Ici, dans le cadre de comparaisons avec NDRAM, qui nécessite un critère de reconstruction parfaite, l'apprentissage bruité pourrait en fait rendre impossible l'atteinte du critère voulu..

### 4.2.4 Étude comparative : Rappel incomplet (stimuli en tons de gris)

Une façon plus réaliste, au niveau cognitif, de mesurer la résistance à la dégradation, est par le retrait de certaines parties de l'image. Ceci serait l'équivalent humain de l'obstruction du champ visuel lorsque l'on observe un objet. Dans la prochaine simulation, plutôt que de retirer des parties complètes d'objets, un ensemble de pixels non-reliés fut retirés. Du point de vue du modèle, ces deux types de dégradation sont toutefois équivalents, puisqu'elles modifient la distance entre le stimulus original et sa version dégradée d'une façon similaire. Ici encore, l'effet de réduction du nombre d'unités de compression fut étudié, et la performance fut une fois de plus comparée avec NDRAM.

#### 4.2.4.1 Méthodologie

Les stimuli, les paramètres et la procédure utilisés pour l'apprentissage sont identiques à ceux de la simulation 4.2.2. Les stimuli ont été sélectionnés car ils permettaient plus de flexibilité que les lettres lorsque vient le temps de changer la valeur des pixels (les lettres ne contiennent que 35 pixels, dont la majorité sont déjà fixés à une valeur de -1, ce qui limite le nombre de niveaux d'incomplétude pouvant être testés). La différence entre la simulation 4.2.2 et celle-ci se situe au niveau du rappel. En effet, lors du rappel, un certain pourcentage de pixels significatifs furent « retirés » de chacun des stimuli. Pour ce faire, lors de la procédure d'incomplétude, les pixels dont la valeur était différente de -1 (fond blanc) furent présélectionnés. De ce nombre (différent pour chaque stimulus), un certain pourcentage prédéterminé fut sélectionné, et la valeur de ces pixels fut fixée à -1. Un exemple des stimuli résultants est présenté à la Figure 4.14, pour 7 niveaux différents d'incomplétude (10% à 70%).

La procédure de rappel fut donc légèrement modifiée, et se détaille comme suit :

1. Essai de rappel :
   a. Sélection de l'un des 10 stimuli d'entrée;
   b. Sélection des pixels faisant partie de l'objet (et non du fond de l'image);
   c. Remplacement d'un certain pourcentage de ces pixels par une valeur de -1;
   d. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $\mathbf{x}(t+1) = \mathbf{x}(t)$);

e. Répétition des étapes *b* à *d* pour 20 rappels incomplets (avec un stimulus incomplet différent pour chaque rappel);
f. Calcul de la proportion de rappels parfaits pour le stimulus;

2. Répétition de l'étape 1 pour chaque stimulus de l'ensemble.

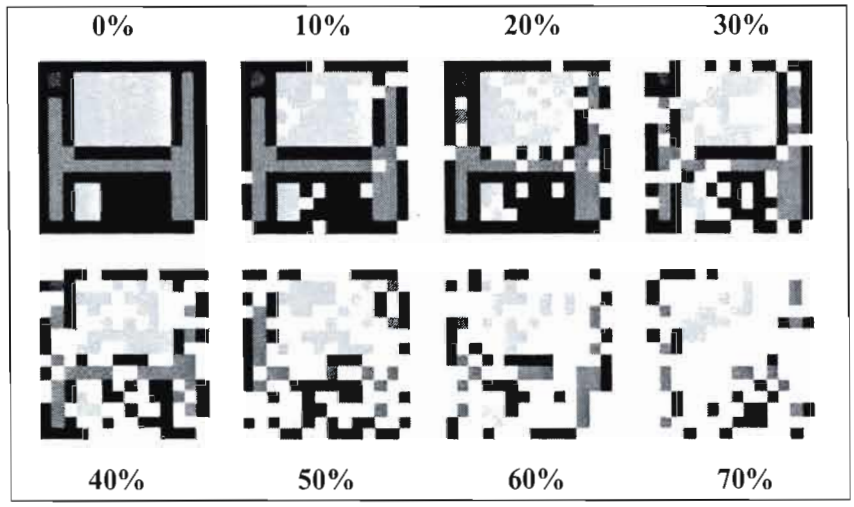


**Figure 4.14** Exemples de stimuli incomplets utilisés pour cette simulation. Le nombre associé à chaque image représente le pourcentage de pixels transformés en pixels blancs (valeur de -1).

Le nombre d'unités de compression (entre 32 et 80, par intervalles de 16), ainsi que le niveau de bruit au rappel (entre 10% et 70%, par intervalles de 10%) furent variés. Pour chaque combinaison de niveau de bruit et de nombre d'unités, 100 simulations, utilisant des matrices de poids aléatoires initiales différentes, furent réalisées (pour un total de 2800 simulations), dans le but d'obtenir une estimation moyenne robuste du pourcentage de rappels parfait. Dans le but de limiter le nombre de simulations à réaliser pour tenir compte des multiples niveaux de chaque variable, le nombre d'unités de compression de FEBAM a été augmenté jusqu'à ce que la performance du réseau soit égale ou supérieure à celle de NDRAM pour tous les niveaux d'incomplétude. Lors du rappel, une reconstruction finale (stable) fut considérée comme acceptable si l'erreur quadratique entre le stimulus d'entrée et la reconstruction était inférieure à $1 \times 10^{-6}$.

4.2.4.2 Résultats

Les résultats sont présentés à la Figure 4.15. Les deux réseaux résistent facilement à un niveau d'incomplétude allant jusqu'à 40% (performance de rappel supérieure à 95%). C'est au-dessus de ce niveau que les différences apparaissent. Ici, à partir d'un niveau de 50%, NDRAM semble nettement désavantagé face à FEBAM, et la différence de performance entre les réseaux s'accroit considérablement à mesure que le niveau d'incomplétude augmente. En effet, le pourcentage de stimuli rappelés parfaitement est supérieur pour FEBAM dans tous les cas, excepté lorsque la compression est réalisée à l'aide de seulement 32 unités. Aussitôt que FEBAM utilise 48 unités de compression ou plus, la performance au rappel est égale ou supérieure à celle de NDRAM.

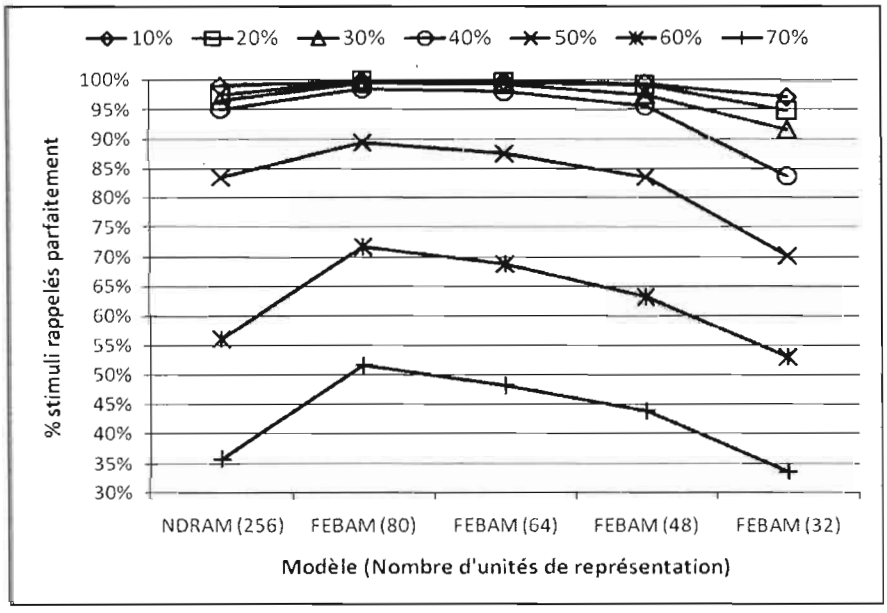


**Figure 4.15** Pourcentage de stimuli en tons de gris rappelés parfaitement en fonction du nombre d'unités de compression et du niveau d'incomplétude. Les résultats obtenus avec NDRAM (100 simulations utilisant les mêmes paramètres qu'avec FEBAM) sont également présentés pour comparaison. Chaque ligne représente un niveau d'incomplétude différent.

Ceci constitue une excellente nouvelle, compte tenu de l'importance supérieure du rappel incomplet (comparativement au rappel avec ajout de bruit) en ce qui a trait à la modélisation de processus humains. FEBAM peut donc mener vers une résistance à ce type

de rappel, tout en permettant une économie de l'ordre de 68.75% (80 unités) à 81.25% (48 unités) au niveau de la taille des représentations (économie de connexions : entre 37.5% et 62.5%).

Une fois de plus, le type de stimuli rappelés diffère entre les deux modèles. FEBAM effectue principalement des erreurs de confusion au rappel, alors que NDRAM se stabilise sur des attracteurs ne faisant pas partie de l'ensemble de départ. Tel que déjà mentionné, l'avantage que FEBAM possède de ce côté pourrait être dû à une résistance accrue aux attracteurs nuisibles, ce qui sera étudié dans la simulation suivante.

### 4.2.5 Étude comparative : Présence d'attracteurs nuisibles (stimuli bipolaires)

Les systèmes récurrents (à base d'attracteurs) possèdent une caractéristique indésirable sous la forme d'attracteurs nuisibles (*spurious states*). Dans les RAM, un attracteur nuisible se définit comme un attracteur ne faisant pas partie de l'ensemble des états (stimuli) de départ. Ainsi, lorsque le réseau effectue un rappel itératif, il est possible qu'il ne se stabilise pas sur l'un des états désirés. La façon classique de déterminer le pourcentage d'attracteurs nuisibles est de faire itérer un ensemble de vecteurs aléatoires dans le réseau, et de vérifier si ces vecteurs convergent vers l'un des états désirés.

Chartier et Proulx (2005) ont comparé NDRAM avec plusieurs autres mémoires autoassociatives récurrentes (Bégin et Proulx, 1996; Diederich et Opper, 1987; Kanter et Sompolinsky, 1987; Storkey et Valabregue, 1999), et ont montré que NDRAM présentait une résistance aux attracteurs nuisibles jusqu'à 8 fois supérieure à ces autres réseaux. FEBAM a pour principal but la reconstruction, et non la stricte stabilité de la représentation à l'apprentissage. Ainsi, les états finaux du réseau devraient plus souvent correspondre aux stimuli d'origine. La prochaine simulation visait donc à vérifier si cette caractéristiques de FEBAM permettaient de réduire le pourcentage d'attracteurs nuisibles, et constituaient donc un avantage marqué.

#### 4.2.5.1 Méthodologie

Les stimuli, les paramètres et la procédure utilisés pour l'apprentissage sont identiques à ceux de la simulation 4.2.1. Une fois de plus, la différence entre la simulation 4.2.1 et celle-ci se situe plutôt au niveau du rappel, qui fut effectué à partir de vecteurs aléatoires. Lorsque la reconstruction finale stable faisait partie de l'ensemble de stimuli originaux, le rappel était considéré comme parfait. Sinon, il fut considéré que le réseau s'était stabilisé sur un attracteur nuisible dans l'espace multidimensionnel.

La procédure de rappel fut donc légèrement modifiée, et se détaille comme suit :

1. Essai d'apprentissage :
   a. Génération aléatoire d'un stimulus-test à l'aide de valeurs contenues dans l'intervalle [-1, 1];
   b. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $\mathbf{x}(t+1) = \mathbf{x}(t)$);
2. Répétition des étapes *a* et *b* pour 500 rappel bruités;
3. Calcul de la proportion de rappels adéquats (reconstruction d'un stimulus faisant partie de l'ensemble de départ).

Le nombre d'unités de compression (entre 11 et 31, par intervalles de 4) fut varié. Pour chaque nombre d'unités, 100 simulations, utilisant des matrices de poids aléatoires initiales différentes, furent réalisées (pour un total de 600 simulations), dans le but d'obtenir une estimation moyenne juste du pourcentage de rappels parfaits. Lors du rappel, une reconstruction finale (stable) fut considérée comme acceptable si l'erreur quadratique entre le cette reconstruction et l'un des stimuli originaux était inférieure à $1 \times 10^{-6}$.

#### 4.2.5.2 Résultats

La Figure 4.16 présente les résultats. On peut voir ici qu'il faut une compression maximale avant que FEBAM atteigne un niveau d'attracteurs nuisibles similaire à celui de NDRAM. La compression représente en fait une augmentation de la charge mnésique du réseau : le nombre de stimuli est le même, mais la taille de l'espace est réduite, ce qui augmente le ratio (ou la charge). La baisse de performance de FEBAM suite à la compression est attribuable à ce facteur : Kanter et Sompolinsky (1987) affirment d'ailleurs que la

proportion d'attracteurs nuisibles augmente en fonction de la charge mnésique. Cela dit, il semble tout de même que le critère d'adéquation du réseau (la reconstruction) ait un impact sur la proportion obtenue de ces attracteurs, qui est inférieure à celle montrée par NDRAM pour la majorité des niveaux de compression.

Il appert donc que l'ajout d'unités dans le réseau, s'il réduit le niveau de compression, permet d'accélérer l'apprentissage, augmenter la résistance au bruit et à l'incomplétude, et réduit considérablement le pourcentage d'attracteurs nuisibles dans l'espace multidimensionnel. La transformation en auto-encodeur d'une RAM permet donc d'augmenter le niveau de polyvalence face aux tâches cognitives, et de préserver (dans la plupart des cas) ou améliorer la performance.

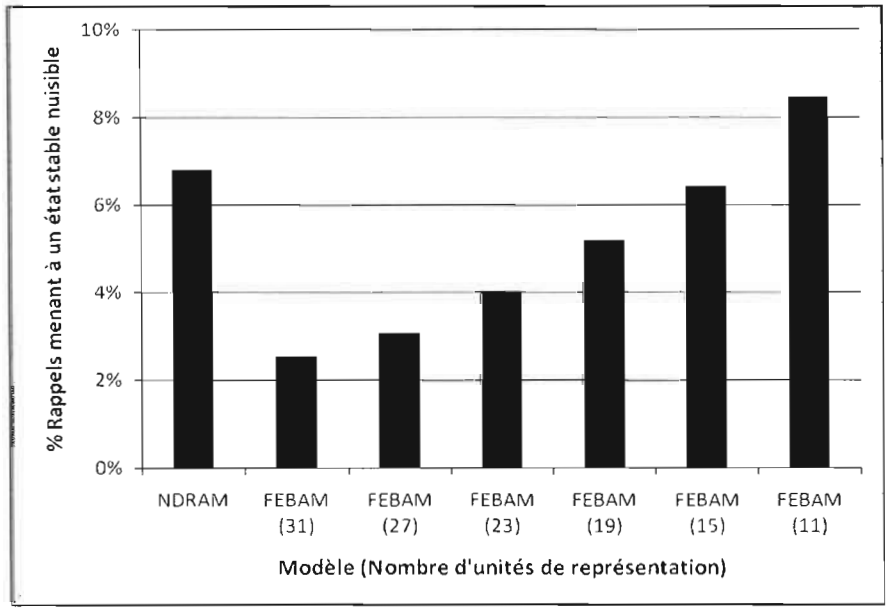


**Figure 4.16** Pourcentage de rappels menant à un attracteur stable nuisible au rappel en fonction du nombre d'unités de compression. Les résultats obtenus avec NDRAM sont également présentés pour comparaison.

## 4.3 Catégorisation et discrimination non-supervisée

De façon générale, la formation de catégories est souvent vue comme une mise en commun de patrons d'entrée similaires dans des catégories communes, un processus s'apparentant aux analyses de type *clustering*, qui impliquent le partitionnement d'espace de

stimuli en un ensemble fini de catégories (ou « grappes »). Dans le domaine des réseaux de neurones, l'analyse de clustering est un principe qui a été développée depuis des années avec les réseaux compétitifs (Kohonen, 1989; Grossberg, 1988). Dans ces modèles, chaque unité de sortie du réseau représente une classe (catégorie). Lorsqu'une décision est prise par le système, l'association entre un exemplaire spécifique et sa « grappe » d'appartenance est renforcée.

Dans les réseaux compétitifs dits « durs » (*hard competitive networks*), tout exemplaire ne peut être associé qu'à une seule catégorie à la fois. Un exemple classique de réseau compétitif dur est celui de la théorie de la résonance adaptive (ART : Grossberg, 1988). Les réseaux ART sont capables de tenir compte du schème « prototype vs. exemplaires », tout en répondant élégamment au dilemme plasticité-stabilité. Cette classe de réseaux compétitifs réussit à reproduire ces comportements désirés grâce à l'ajout d'un détecteur de nouveauté (à travers un principe de « vigilance »); ainsi, plusieurs niveaux de généralisation peuvent être atteints à l'aide d'une même procédure. Si la valeur du paramètre de vigilance est basse, des catégories vastes seront développées, alors que si elle est élevée, plus de catégories étroites se développeront, et le réseau effectuera ultimement de l'apprentissage d'exemplaires.

En « soft computing » (Kohonen, 1989), chacun des exemplaires peut être associé à différentes « grappes », et ce, à des degrés différents. Ce principe permet d'enregistrer des classifications plus distribuées; par exemple, un exemplaire pourrait être géométriquement situé entre deux grappes, et posséder ainsi divers degrés d'appartenance catégorielle.

Les réseaux à base de PCA (Diamantaras et Kung, 1996) peuvent aussi être utilisés pour effectuer de l'analyse en grappes. Dans ce cas, chaque catégorie est définie par une somme linéaire de composantes orthogonales. Les réseaux de PCA non-linéaires (Karhunen, Pajunen et Oja, 1998) ne sont pas limités par cette exigence orthogonale; ainsi, des composantes corrélées, servant à définir les catégories perceptuelles, peuvent être extraites.

Dans tous les cas présentés, lorsqu'un item est associé à une grappe (catégorie) spécifique, il n'existe aucun mécanisme permettant de modifier l'appartenance catégorielle. La structure interne de chacun de ces modèles est basée sur une métrique spécifique qui est constante au long de la période d'apprentissage. Dans les réseaux compétitifs, l'association

entre un stimulus et la ou les unités gagnantes ne peut qu'être renforcée durant l'apprentissage. En général, aucun mécanisme ne permet l'affaiblissement ou la modification des relations. Dans les réseaux à base de PCA, chaque composante développée est fixée suite à un processus de convergence séquentiel, et chaque composante subséquente doit par définition être orthogonale ou indépendante des précédentes. Il serait donc impossible de modifier l'ensemble de composantes sans recommencer le processus entier et laisser de côté la contribution informative des apprentissages précédents.

Avec FEBAM, le type d'apprentissage utilisé (hebbien/anti-hebbien, sans connexions inhibitrices latérales) devrait permettre au réseau de réorganiser les catégories durant la procédure d'apprentissage, tout comme le font les humains (voir Murphy, 2002). Ceci est possible seulement dans le cas où l'on ajoute des unités de compression dans le réseau à un ou des moments choisis. Lorsque l'on rajoute des unités dans la couche **y**, on agrandit par le fait même l'espace des représentations en y ajoutant une dimension spatiale. Lorsque cet espace multidimensionnel est agrandi, cela permet de développer un plus grand nombre de divisions inter-catégorielles, et par conséquent, un plus grand nombre de catégories. Ainsi, il est attendu que le paysage catégoriel devrait demeurer relativement stable à travers le temps, mais que certains stimuli pourrait être catégorisés avec des membres différents, à mesure que l'apprentissage se complète. Aussi, il est attendu que la catégorisation devrait se faire de plus en plus précise à mesure que la taille de l'espace augmente, et que des séparations plus fines deviennent possibles.

### 4.3.1 Étude de comportement : Rappel selon le nombre d'unités de compression

Cette simulation vise principalement à déterminer la relation entre le nombre d'unités de compression dans le réseau, et le nombre de catégories développées, lorsque les stimuli ont une appartenance catégorielle prédéterminée. Si le réseau réussit à augmenter le nombre de catégories en faisant usage d'un espace vectoriel agrandi (plus d'unités de compression), il serait alors sensé de tester la validité d'un possible mécanisme itératif d'ajout d'unités dans le réseau.

4.3.1.1 Méthodologie

Quatre catégories, bâties autour de prototypes, furent créées pour cette simulation. Chaque vecteur prototypique, d'une taille de 100 positions, fut généré aléatoirement, avec la contrainte que la corrélation moyenne entre les prototypes devait être supérieure à $r = 0.1$ mais inférieure à $r = 0.2$. Les quatre prototypes utilisés sont illustrés à la Figure 4.17. Pour chacune des catégories, 10 exemplaires furent créés en inversant la valeur de 6 pixels à partir du prototype. Ainsi, la corrélation entre chaque exemplaire et son prototype d'origine était égale à $r = 0.88$, et les corrélations intra-catégorielles (entre exemplaires d'une même catégorie) moyennes étaient égales à $r = 0.76$. Chaque ensemble de pixels inversés était exclusif. La Figure 4.18 présente quatre exemplaires associés à un prototype. Les corrélations inter-prototypes sont présentées au Tableau 4.1. La corrélation moyenne inter-prototype (en valeur absolue) était égale à $r = 0.12$. Il est à noter que seuls les 40 exemplaires furent présentés à l'apprentissage et au rappel; les prototypes d'origine ne furent jamais présentés.

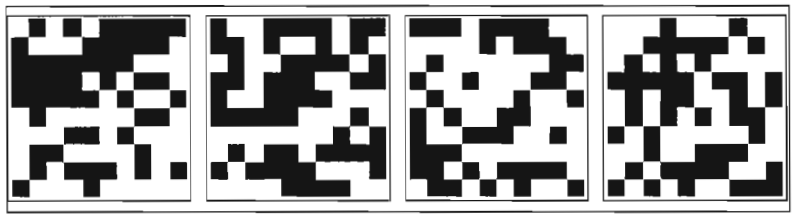

**Figure 4.17** Prototypes ayant servi à générer les quatre catégories.

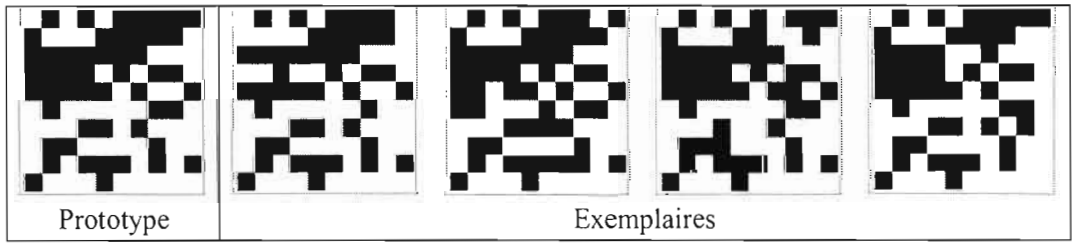


**Figure 4.18** Quatre exemplaires générés à partir d'un prototype.

**Tableau 4.1** Corrélations inter-prototypes ($r$)

| | *Prototype 2* | *Prototype 3* | *Prototype 4* |
|---|---|---|---|
| *Prototype 1* | 0.08 | 0.08 | 0.22 |
| *Prototype 2* | | -0.19 | -0.10 |
| *Prototype 3* | | | -0.02 |

Les paramètres de FEBAM furent fixés aux valeurs suivantes : le paramètre général de transmission fut fixé à $\delta = 0.1$ et le paramètre général d'apprentissage fut fixé à $\eta = 0.0025$ (valeur maximale : 0.006). La procédure complète se déroula comme suit :

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]);
1. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le réseau, tel qu'illustré à la Figure 3.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   a. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 3.3 et 3.4;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Calcul de l'erreur quadratique moyenne pour l'ensemble de stimuli;
4. Répétition des étapes 1 à 3 jusqu'à ce que la différence d'erreur quadratique moyenne entre deux blocs consécutifs d'apprentissage soit inférieure ou égale à $1 \times 10^{-6}$.

La procédure de rappel se déroula comme suit :

1. Essai de rappel :
   a. Sélection d'un stimulus de l'ensemble d'entrée;
   b. Itération dans le réseau (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une compression stable (en l'occurrence, lorsque $\mathbf{y}(t+1) = \mathbf{y}(t)$);
2. Répétition de l'étape 1 pour chaque stimulus de l'ensemble;
3. Détermination du nombre de catégories dans le paysage catégoriel (si deux stimuli convergent vers la même compression stable, ils font partie de la même catégorie);

Le nombre d'unités de compression dans le réseau fut varié (entre 1 et 30 unités). Pour chaque nombre d'unités, 50 simulations, utilisant des matrices de poids aléatoires initiales différentes, furent réalisées (pour un total de 1500 simulations), dans le but d'obtenir une estimation robuste du nombre de catégories développées par le réseau.

4.3.1.2 Résultats

Les résultats sont présentés à la Figure 4.19. On peut clairement voir que le nombre d'unités est relié de façon croissante et monotone au nombre de catégories moyen dans le paysage catégoriel. Ainsi, FEBAM peut passer d'un processus de catégorisation (plusieurs items par « grappe ») à un processus d'identification (un item par « grappe ») sans aucune modification aux postulats de base, à l'architecture et aux règles.

Au début de l'apprentissage, le nombre de catégories n'augmente pas automatiquement aussitôt que l'on ajoute une unité de compression. Ceci mène à croire qu'un niveau minimal d'apprentissage perceptuel (extraction de composantes) est nécessaire avant qu'une différentiation catégorielle puisse se faire. Avant ce niveau minimal, une confusion inter-stimuli persiste. À partir de ce niveau minimal de reconstruction de l'entrée, l'ajout d'unités a un effet direct sur le nombre de catégories, puisque le réseau possède suffisamment d'informations pour différencier divers groupes de stimuli. À la fin de l'apprentissage, l'ajout d'unités a également un effet moindre. Ceci est causé par la structure des catégories, car les corrélations entre les stimuli d'une même catégorie sont plus élevées que celles entre ces stimuli et les prototypes des autres catégories. Ceci rend la différentiation entre stimuli plus difficile. En effet, pour totalement différencier les stimuli (processus d'identification), le réseau doit développer des composantes de plus en plus précises, idiosyncratiques, qui tiennent compte de différences inter-stimuli minimes. C'est pourquoi l'ajout d'une dimension à l'espace représentationnel n'a pas nécessairement le même effet qu'au milieu de l'apprentissage.

D'un point de vue théorique, ce résultat est extrêmement important, puisqu'il permet de conceptualiser la différence entre catégorisation et identification non pas comme qualitative, mais plutôt quantitative. Les deux phénomènes constitueraient ici les deux extrêmes d'un continuum où la seule différence est le nombre d'unités de compression utilisées. Ceci simplifie le nombre de postulats nécessaires pour modéliser les deux tâches (plusieurs auteurs, tels Nosofsky (1988), malgré la reconnaissance d'un lien entre les deux processus, postulent des mécanismes qualitativement distincts).

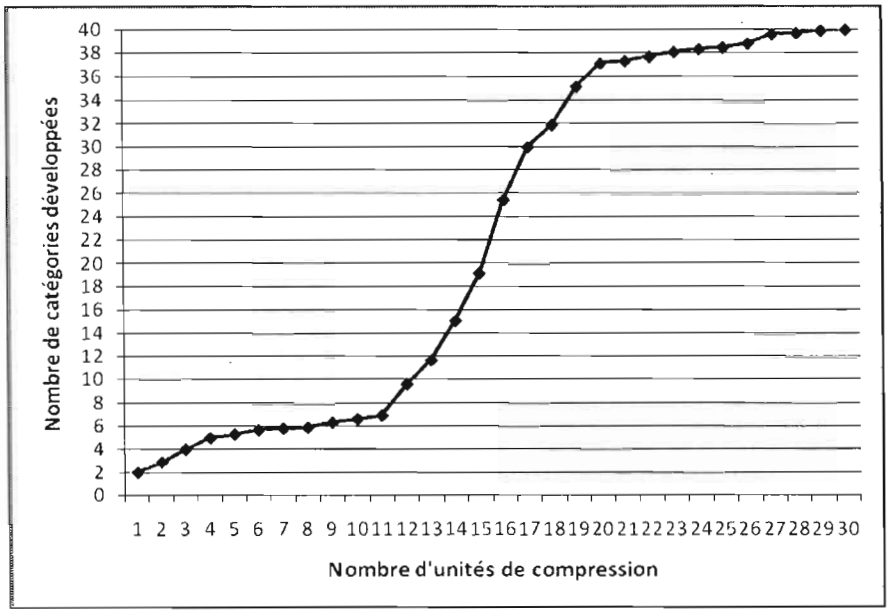


**Figure 4.19** Nombre de catégories développées en fonction du nombre d'unités dans le réseau. Le même ensemble de stimuli fut utilisé pour toutes les simulations.

Ainsi, si l'on désirait ultimement modéliser la double dissociation (Glanzer et Cunitz, 1966) cognitive et neurologique mise en évidence par Knowlton et Squire (1993; Knowlton, Mangels et Squire, 1996; Marsolek, 1995; Reber, Stark et Squire, 1998a, 1998b; Squire et Knowlton, 1995), les deux systèmes postulés pourraient être des réseaux FEBAM dont le nombre d'unités de compression diffère. Le réseau permettant la mémoire épisodique d'exemplaires par processus d'identification contiendrait un plus grand nombre d'unités, et le réseau effectuant la catégorisation par abstraction en contiendrait un minimum.

### 4.3.2 Étude de la nécessité d'un processus adaptif d'ajout d'unités

Maintenant qu'il a été montré que FEBAM peut passer d'un processus de catégorisation à un processus d'identification lorsque le nombre d'unités croît, il serait intéressant d'étudier le phénomène non pas à travers un ensemble de simulations indépendantes moyennées, mais plutôt à l'intérieur d'un unique processus d'apprentissage et de rappel.

4.3.2.1 Méthodologie

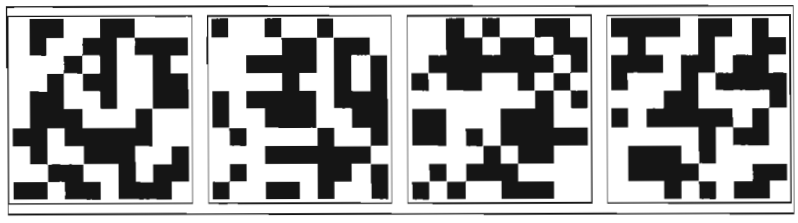

**Figure 4.20** Prototypes ayant servi à générer les quatre catégories.

**Tableau 4.2** Corrélations inter-prototypes ($r$)

| | *Prototype 2* | *Prototype 3* | *Prototype 4* |
|---|---|---|---|
| *Prototype 1* | 0.02 | -0.10 | 0.04 |
| *Prototype 2* | | -0.12 | 0.02 |
| *Prototype 3* | | | -0.02 |

Quatre nouvelles catégories, bâties autour de prototypes furent créées en utilisant la même procédure que pour la simulation précédente. Pour encourager la séparation catégorielle dès le début du processus d'apprentissage, la contrainte de corrélation moyenne entre les prototypes fut abaissée (minimum $r$ = 0.05, maximum : $r$ = 0.1). Les quatre prototypes utilisés sont illustrés à la Figure 4.20. Chaque catégorie contenait 10 exemplaires. Les corrélations inter-prototypes sont présentées au Tableau 4.2. La corrélation moyenne inter-prototype (en valeur absolue) était égale à $r$ = 0.05. Les paramètres de FEBAM furent fixés aux mêmes valeurs que pour l'ensemble de simulations précédent. La procédure complète se déroula comme suit:

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]);
1. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le réseau, tel qu'illustré à la Figure 3.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 3.3 et 3.4;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Calcul de l'erreur quadratique moyenne pour l'ensemble de stimuli;

4. Répétition des étapes 1 à 3 jusqu'à ce que la différence d'erreur quadratique moyenne entre deux blocs consécutifs d'apprentissage soit inférieure ou égale à $1x10^{-6}$.
5. Essai de rappel :
   a. Sélection d'un stimulus de l'ensemble d'entrée;
   b. Itération dans le réseau (à l'aide des matrices W et V finales) jusqu'à l'atteinte d'une reconstruction stable (en l'occurrence, lorsque $\mathbf{y}(t+1) = \mathbf{y}(t)$);
6. Répétition de l'étape 5 pour chaque stimulus de l'ensemble;
7. Détermination du nombre de catégories dans le paysage catégoriel (si deux stimuli convergent vers la même compression stable, ils font partie de la même catégorie);
8. Ajout d'une unité dans la couche de compression **y**;
9. Initialisation aléatoire des poids de connexion pour la nouvelle colonne ajoutée à la matrice **W** et la nouvelle ligne ajoutée à la matrice **V** (Intervalle des valeurs de départ : [-0.1, 0.1]);
10. Répétition des étapes 1 à 9 jusqu'à ce que le nombre de catégories soit égal au nombre de stimuli inclus dans l'ensemble de départ.

#### 4.3.2.2 Résultats

La Figure 4.21 présente l'erreur quadratique en fonction du nombre de blocs d'apprentissage complétés. On peut voir que chaque ajout d'unité permet de réduire l'erreur quadratique, mais que l'effet de ces ajouts est de moins en moins marqué à mesure que le nombre d'unités augmente. La Figure 4.22 valide le processus d'ajout d'unités dans le réseau. En effet, lorsqu'il n'y a qu'une unité dans le réseau, ce dernier crée deux catégories, et à mesure que des unités sont ajoutées, le nombre de catégories augmente (ici, de façon quasi-monotone[40]), jusqu'à ce que 40 catégories (une catégorie par exemplaire : processus d'identification) soient présentes dans le paysage multidimensionnel. On retrouve ici le même effet qu'à la simulation précédente. L'ajout d'unités a un effet maximal sur le nombre de catégories vers le milieu de la procédure.

Tel que le montre la Figure 4.23, aussitôt que le réseau contient 3 unités, la représentation catégorielle devient fidèle à l'appartenance prédéfinie (ce qui est tout de même prévisible, vu les faibles corrélations inter-prototypes, mais valide néanmoins l'utilité du réseau pour une telle tâche). Bien que l'étude du paysage catégoriel montre que certains items changent de « partenaires » catégoriels lors d'ajouts d'unités, les membres d'une même

[40] Il est à noter qu'il ne s'agit que d'une seule simulation; l'étude précédente a montré qu'en moyenne, le lien entre le nombre d'unités et le nombre de catégories s'accroissait de façon monotone.

catégorie sont classés ensemble. Le phénomène de réorganisation sera plus évident à la simulation suivante.

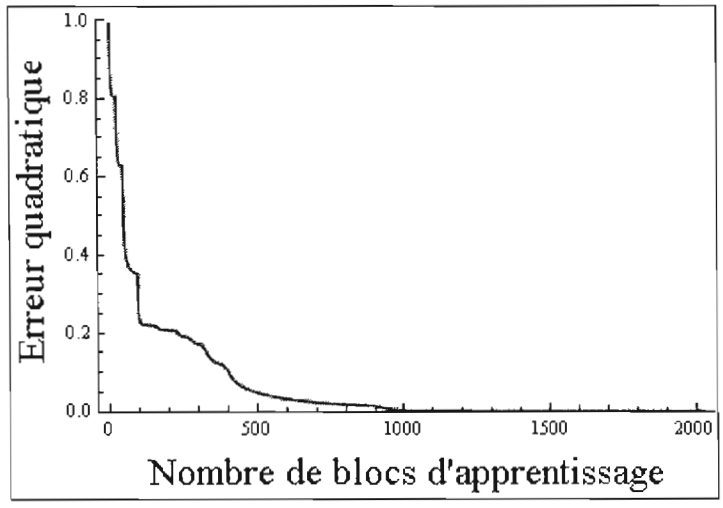


**Figure 4.21** Erreur quadratique moyenne en fonction du bloc d'apprentissage.

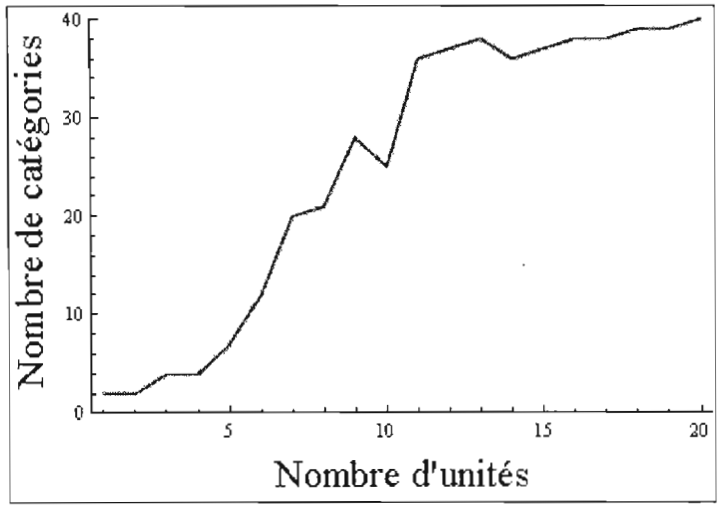


**Figure 4.22** Nombre de catégories développées par FEBAM en fonction du nombre d'unités de compression.

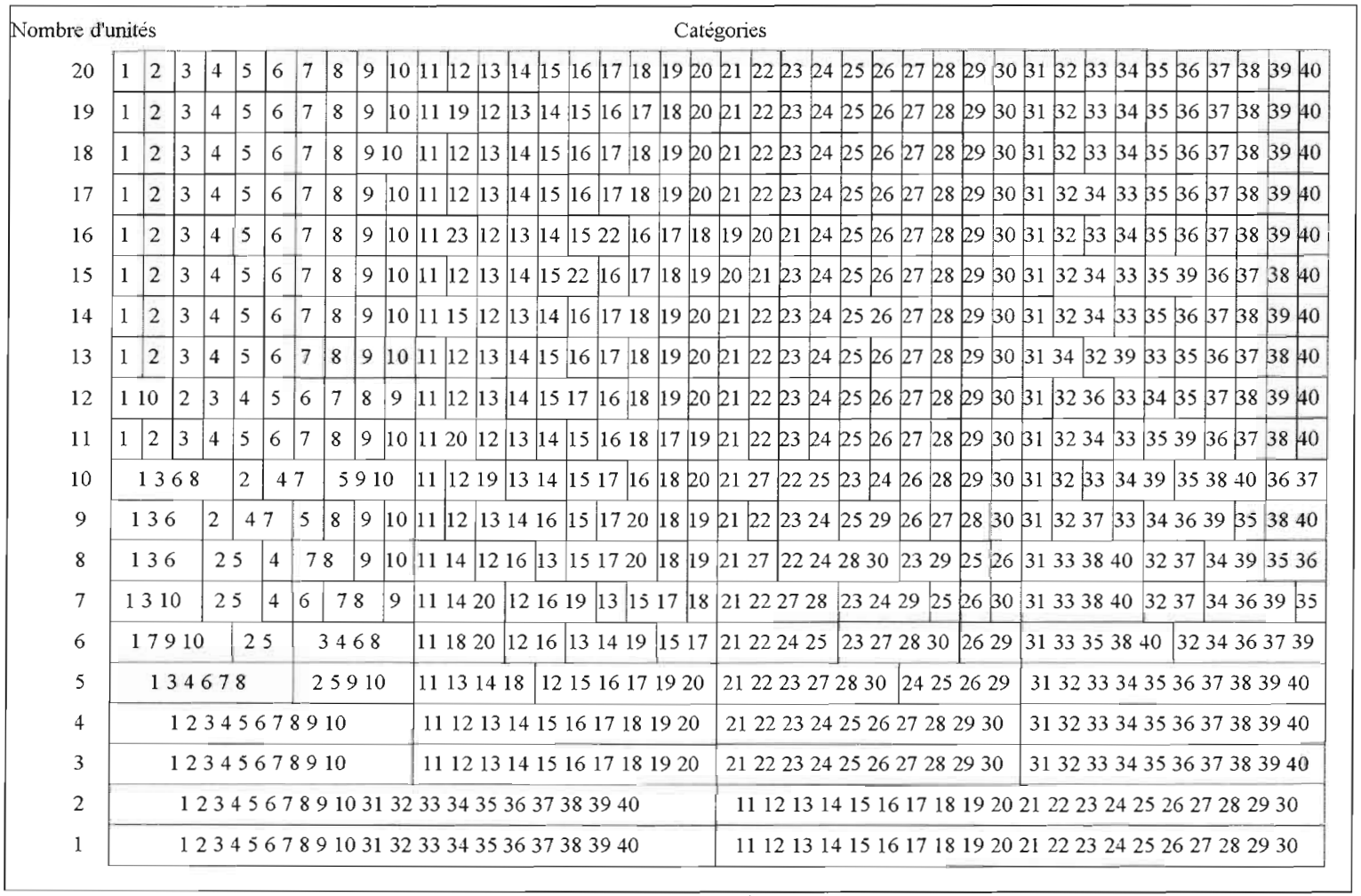


**Figure 4.23** Catégories développées par le réseau suite à l'ajout de chacune des unités de compression supplémentaires.

Même si le critère d'ajout d'unités (arrêt de l'amélioration de l'erreur quadratique) est peu plausible cognitivement, cette simulation démontre une fois de plus la polyvalence du réseau, et surtout, le possible caractère quantitatif de la différence entre catégorisation et identification.

### 4.3.3 Réplication : Appartenance catégorielle indéfinie

#### 4.3.3.1 Méthodologie

**Figure 4.24** Stimuli utilisés pour cette simulation.

Les stimuli utilisés pour cette simulation sont présentés à la Figure 4.24. Il s'agit de représentations bipolaires de lettres majuscules de l'alphabet, en format 7 pixels par 7 pixels. Ces stimuli ont été choisis pour leur absence d'appartenance catégorielle prédéterminée, et parce qu'ils présentent un large éventail de corrélations inter-stimuli (entre $r = 0.01$ et $r = 0.84$).

Les paramètres de FEBAM furent fixés aux valeurs suivantes : le paramètre général de transmission fut fixé à $\delta = 0.1$, et le paramètre général d'apprentissage fut fixé à $\eta = 0.0025$ (valeur maximale: 0.012). La procédure utilisée fut identique à celle de la simulation précédente.

#### 4.3.3.2 Résultats

Les Figures 4.25 et 4.26 répliquent les résultats obtenus dans la simulation précédente. Ici, un résultat important est que seules 15 unités de compression sont nécessaires pour discriminer les stimuli de l'ensemble. Lors d'une simulation de rappel parfait effectué avec les mêmes stimuli, il a été déterminé qu'un minimum de 40 unités de compression était nécessaire pour une reconstruction parfaite (Giguère, Chartier, Proulx et Lina, 2007a). Ainsi, le critère moins

strict de discrimination/différentiation permet de croire que l'économie cognitive est encore plus marquée que lorsque le rappel parfait est visé, alors que le critère visé ici est beaucoup plus souple, mais amplement suffisant pour la tâche.

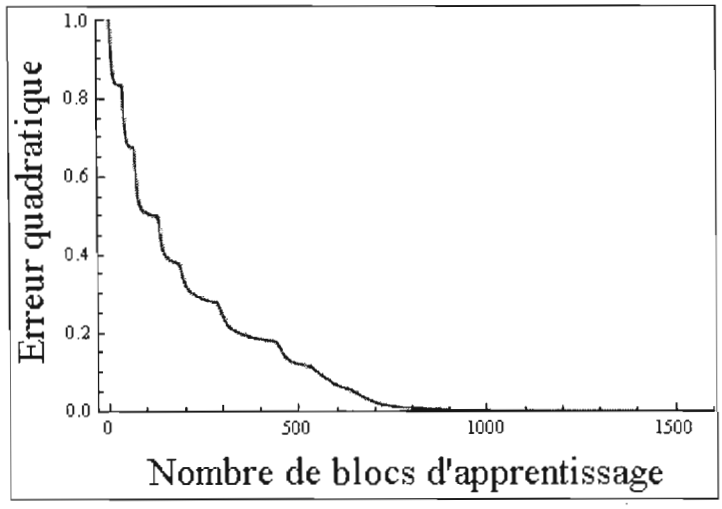


**Figure 4.25** Erreur quadratique moyenne en fonction du bloc d'apprentissage.

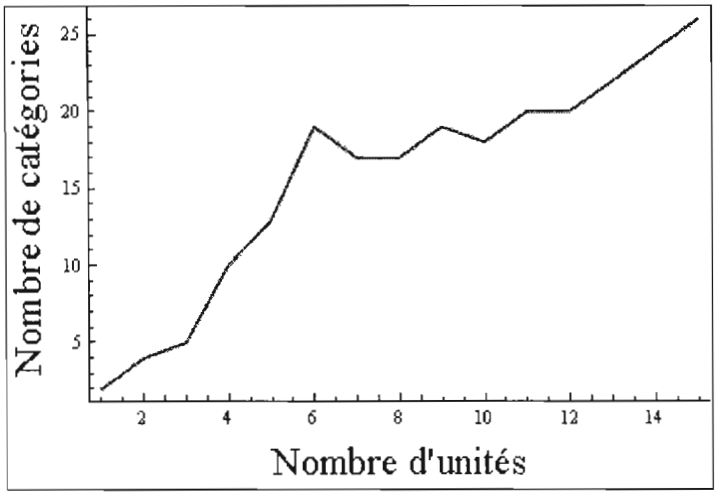


**Figure 4.26** Nombre de catégories développées par FEBAM en fonction du nombre d'unités de compression.

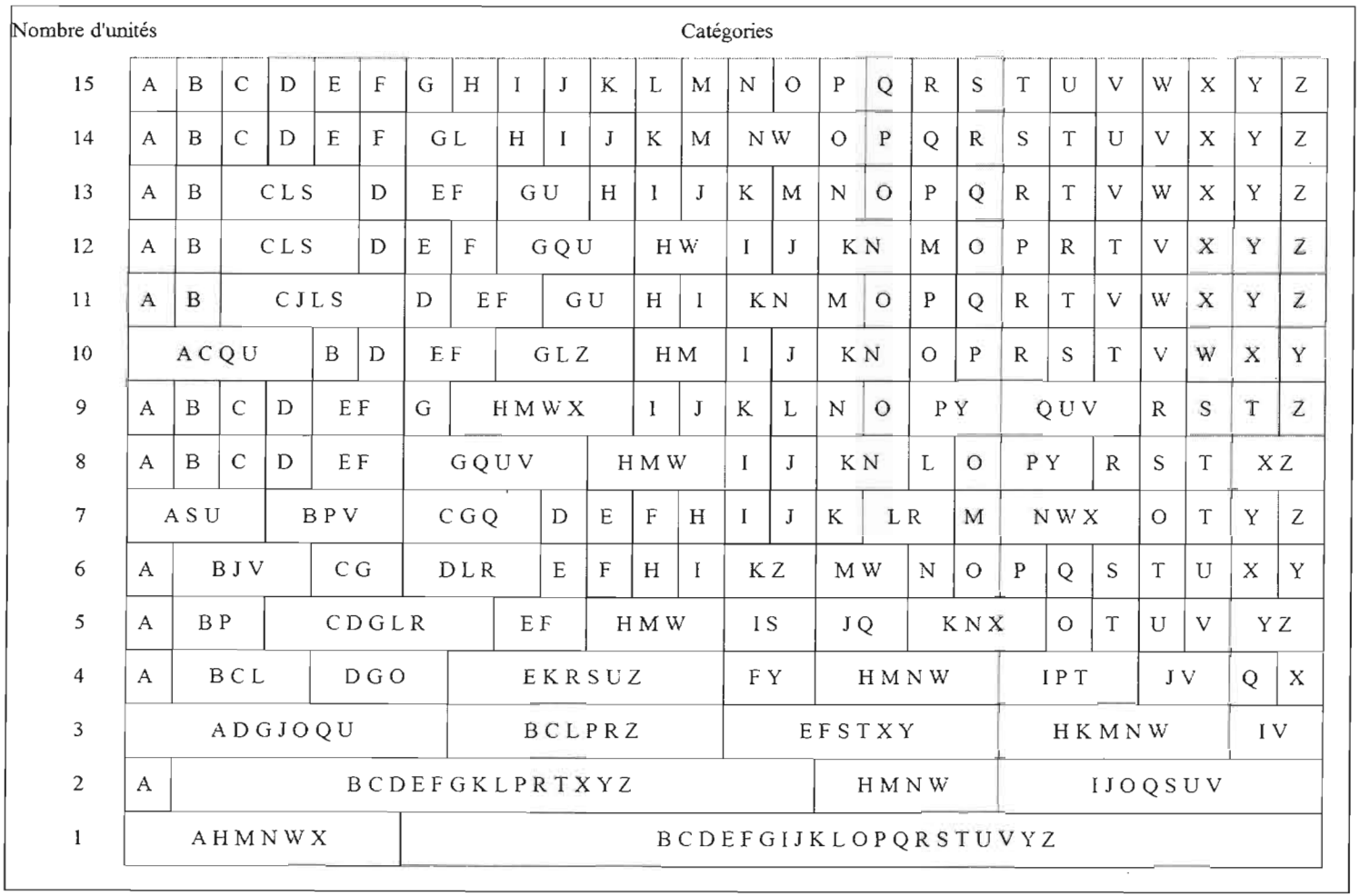


**Figure 4.27** Catégories développées par le réseau suite à l'ajout de chacune des unités de compression supplémentaires.

Un autre résultat intéressant est illustré à la Figure 4.27. Ici, l'absence d'appartenance catégorielle prédéterminée nous permet de mieux apprécier la réorganisation des catégories. Ainsi, tout dépendant de l'information pertinente, la lettre I est associée avec 19 autres lettres lorsqu'il n'y a qu'une unité, pour ensuite demeurer associée à un sous-ensemble de ces lettres à 2 unités. Ensuite, cette lettre forme une catégorie avec la lettre V, et suite à l'ajout d'une quatrième unité, est maintenant pairée avec P et T, et ensuite avec S, avant de former sa propre catégorie. D'autres stimuli, tels les lettres E et O, ont un parcours plus stable, et sont constamment pairées avec les mêmes stimuli, ce qui amène un minimum de stabilité au processus. Le réseau est donc sensible aux nouvelles informations diagnostiques, et en tient compte dans la création itérative catégorielle, tout en permettant une certaine stabilité.

## 4.4 Conclusion

Les nombreuses simulations présentées dans ce chapitre ont permis d'établir la polyvalence de FEBAM en ce qui a trait aux processus d'apprentissage et de catégorisation perceptuels non-supervisés. Ces simulations ont principalement montré l'utilité du double processus de compression en représentation économique et d'extraction de caractéristiques en vue d'une reconstruction de plus en plus précise du patron d'entrée.

En plus de montrer la validité du processus d'extraction de composantes, il a été montré que FEBAM permet des économies cognitives au niveau de la taille des représentations et du nombre de connexions devant être mises à jour dans le réseau. Il existe une relation négative marquée entre le niveau d'économie cognitive et certaines variables telles la rapidité d'apprentissage et la tolérance au bruit, mais seulement lorsque la compression est maximale.

Suite à de multiples comparaisons avec le modèle duquel FEBAM émerge (soit NDRAM), on conclut que ce dernier possède une tolérance à l'incomplétude supérieure, une propension moindre à créer des attracteurs nuisibles lors de l'apprentissage, et une possibilité de modélisation plus réaliste du type d'erreur possible au rappel, soit la confusion entre deux stimuli existants.

Finalement, il a été montré que ce même réseau peut effectuer des tâches de mise en commun d'exemplaires grâce au processus de compression. La catégorisation et

l'identification peuvent être vues comme un seul processus, qui ne diffère que quantitativement.

Il est pertinent de rappeler ici que les tâches réalisées par le réseau constituent des propriétés intrinsèques de celui-ci; en effet, toutes ont été effectuées à l'aide d'un seul ensemble de postulats théoriques, d'une seule architecture, une seule règle d'apprentissage et une seule règle de transmission, ce qui rend le modèle extrêmement parcimonieux.

# CHAPITRE V

## AJOUT D'UN MODULE D'ASSOCIATION DE RÉPONSE

Un grand nombre de modèles symboliques de catégorisation (par exemple, Nosofsky, 1986;1988;1992; Kruschke, 1992) postulent que l'apprentissage catégoriel implique le développement d'un ensemble de poids attentionnels. La valeur de ces poids indique la diagnosticité perçue de chacune des caractéristiques pour une catégorie donnée. Le coefficient d'attention représente une force d'association entre une caractéristique symbolique et une étiquette catégorielle.

Dans le cas de FEBAM, les caractéristiques développées ne sont pas symboliques, mais chaque entrée perceptuelle est tout de même définie de façon comprimée à l'intérieur du système. On pourrait donc associer chaque version comprimée d'une entrée perceptuelle avec un vecteur de réponse prédéterminé. Chaque poids de connexion entre une couche de compression et une couche de réponse représenterait alors la diagnosticité d'une composante développée en lien avec l'étiquette apprise.

Dans ce chapitre, une version supervisée (*i.e.* tenant compte de réponses désirées) de FEBAM sera proposée. Pour des raisons de consistance interne, ce second modèle doit conserver les principes définitoires du premier: apprentissage hebbien/anti-hebbien, même règle de transmission, même règle d'apprentissage. La seule différence est dans l'architecture, où un module supplémentaire est inséré afin que le modèle puisse avoir accès à un superviseur externe. Ce réseau sera donc en mesure de modéliser un apprentissage supervisé, sans devoir proposer un modèle hybride (contenant des classes de réseaux différentes), et surtout sans devoir augmenter le nombre de postulats de base liés à l'approche.

## 5.1 Description du modèle

### 5.1.1 Architecture

Le modèle FEBAM-RA (RA pour *Response Association* : Figure 5.1) contient deux modules. Le premier module sert au traitement perceptuel, et est lié directement à l'environnement. C'est ce module qui reçoit les entrées initiales **x**(0). Il s'agit en fait d'un réseau FEBAM (Figure 3.1), tel que décrit à la section 3.1. Le second module est une BHM (Chartier et Boukadoum, 2006) servant à associer la représentation comprimée du stimulus **y**(1) avec une étiquette identificative ou catégorielle (fournie de façon externe). Ceci est analogue à une rétroaction reçue lors d'une tâche de laboratoire en psychologie cognitive. On nommera donc ce module le « module d'association de réponse ». L'architecture de ce second module est illustrée à la Figure 5.1. L'utilisation conjointe de FEBAM et d'une BHM permet de réduire au minimum les postulats théoriques et techniques. FEBAM, tout comme NDRAM, constitue un cas spécial de la BHM (sans entrée externe pour la seconde couche d'unités). Ainsi, l'architecture des deux modules sera quasi-identique. Les deux modules partageront aussi le principe d'apprentissage hebbien/anti-hebbien basé sur les différences temporelles. Finalement, les règles d'apprentissage et de transmission des modules suivront les mêmes formes. Ceci constitue donc une combinaison nettement parcimonieuse.

Le réseau FEBAM-RA fonctionnera en deux étapes. La première étape sera équivalente au traitement effectué par le réseau FEBAM, soit un cycle tel que défini à la Figure 3.2. Ce premier module utilisera donc des entrées perceptuelles provenant de l'environnement pour créer des représentations internes comprimées. Une fois que ce traitement perceptuel sera effectué, la compression créée pour le stimulus sera associée à un vecteur représentant l'appartenance de ce stimulus. Cette « étiquette » d'appartenance pourra servir à l'identification (lorsque chaque vecteur distinct est lié à une seule entrée perceptuelle) ou à la catégorisation (lorsque chaque vecteur distinct est simultanément lié à plusieurs entrées perceptuelles). Le principe d'association d'une étiquette de réponse simplifie grandement le réel processus décisionnel effectué lors d'une tâche de laboratoire. Ceci est effectué dans le but de garder, pour l'instant, le nombre de principes définitoires au minimum. Cette façon de modéliser la rétroaction externe est inspirée de Gerganov, Grinberg, Quinn

et Goldstone (2007), qui ont cependant associé les étiquettes catégorielles aux stimuli de façon unidirectionnelle (par opposition à bidirectionnelle dans FEBAM-RA).

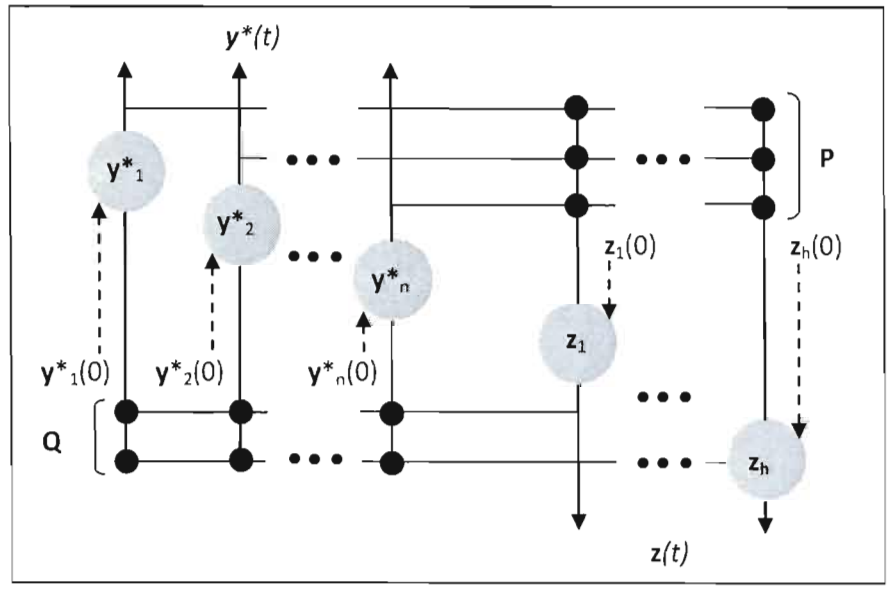


**Figure 5.1** Architecture du module d'association de réponse. Ce module est en fait une BHM (Chartier et Boukadoum, 2006). On y associe la sortie du module perceptuel (le vecteur **y**(*t*), qui devient ici **y***(0) lorsqu'il est copié dans le deuxième module) au vecteur-réponse prédéterminé **z**(0). Le réseau contient donc deux couches d'unités, et chaque unité de la couche **y*** est connectée à chaque unité de la couche **z**. Il n'y aucune connexion entre unités d'une même couche. Les lignes pointillées indiquent les deux points d'entrée du réseau.

La Figure 5.2 illustre le processus suivi par FEBAM-RA lors de la présentation d'un stimulus à l'apprentissage. Pour débuter, le module perceptuel transforme l'entrée à dimensionnalité complète **x**(0) en une version réduite ou comprimée d'elle-même **y**(0), à l'aide du produit entre la matrice **W** et le vecteur d'entrée. Ensuite, cette version comprimée passe à travers la matrice **V**, dans le but de produire une reconstruction finale **x**(1). Cette reconstruction passe une deuxième fois par la matrice **W**, produisant la compression finale **y**(1). Cette compression finale est copiée (*transmission identitaire*) dans la couche d'entrée **y*** du module d'association de réponse; on note cette copie de compression **y***(0). Tout comme avec la BHM, le réseau transformera la copie de compression **y***(0) en vecteur de réponse **z**(1), à l'aide du produit matriciel avec **P**. De façon simultanée, le module transformera la réponse désirée **z**(0) en compression **y***(1), à l'aide de la matrice **Q**. C'est ce

double processus simultané qui produit l'association bidirectionnelle entre une compression et une réponse.

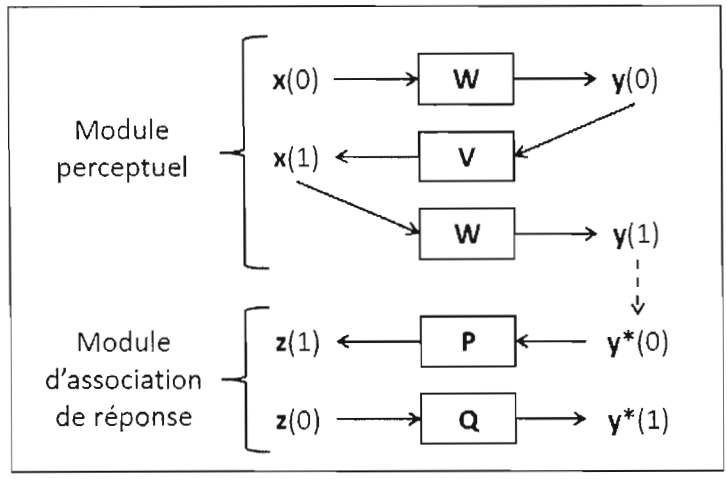


**Figure 5.2** Schéma illustratif du processus itératif (ou *cycle*) réalisé par le réseau avant chacune des mises à jour des matrices de poids de connexion. La flèche pointillée représente ici une simple copie (un transfert direct) du contenu d'une couche à une autre. Le module perceptuel est un réseau FEBAM et le module d'association de réponse est une BHM. Dans la présente thèse, le nombre de cycles réalisés avant la mise à jour des poids de connexion dans chaque module sera toujours égal à 1.

### 5.1.2 Règle de transmission

La règle de transmission utilisée par le module perceptuel est la même qu'aux équations 3.1 et 3.2 (Section 3.1.2). Par parcimonie, cette même règle est également utilisée pour le module d'association de réponse. Pour ce dernier, la règle adaptée devient donc :

$$\forall i,\ldots,H, \mathbf{z}_i(t+1) = \begin{cases} 1, \ Si\ \mathbf{Qy}_i^*(t) > 1 \\ -1, \ Si\ \mathbf{Qy}_i^*(t) < -1 \\ (\delta+1)\mathbf{Qy}_i^*(t) - \delta(\mathbf{Qy}^*)_i^3(t), \ Sinon \end{cases} \quad 5.1$$

$$\forall i,\ldots,N, \mathbf{y}_i^*(t+1) = \begin{cases} 1, \ Si\ \mathbf{Pz}_i(t) > 1 \\ -1, \ Si\ \mathbf{Pz}_i(t) < -1 \\ (\delta+1)\mathbf{Pz}_i(t) - \delta(\mathbf{Pz})_i^3(t), \ Sinon \end{cases} \quad 5.2$$

où $H$ représente le nombre d'unités dans la couche de réponse associée $\mathbf{z}$, $N$ représente le nombre d'unités dans la copie de la couche de compression $\mathbf{y}^*$, $i$ représente l'indice de

l'élément du vecteur respectif, $\mathbf{z}(t)$ et $\mathbf{y}^*(t)$ représentent les contenus des couches d'unités au temps $t$, **Q** et **P** représentent les matrices de poids de connexion asymétriques liant les deux couches, et $\delta$ est un paramètre général de transmission (dont la valeur sera fixée à 0.1 pour les deux modules). La matrice **Q** mémorise la fonction transformant les compressions en réponses. La matrice **P** permet de récupérer la compression associée à une étiquette identificative ou catégorielle. Dans le cas d'une étiquette catégorielle, par le biais des propriétés du BHM, le vecteur récupéré sera le prototype de la catégorie[41].

### 5.1.3 Règle d'apprentissage

La règle d'apprentissage utilisée par le module perceptuel est la même qu'aux équations 3.3 et 3.4 (Section 3.1.3). Pour le modèle FEBAM-RA, deux paramètres d'apprentissage distincts seront utilisés. Le paramètre d'apprentissage du module perceptuel sera maintenant représenté par le symbole $\eta_P$, alors que celui du module d'association de réponses sera identifié par le symbole $\eta_A$. La distinction entre ces deux valeurs est nécessaire pour le respect de la condition de stabilité énoncée à l'équation 3.5. Cette condition est calculée en fonction de la taille maximale des couches d'un module. Par définition, la taille de la couche avec le plus d'unités dans le module perceptuel (*i.e.* la couche **x**) sera plus élevée que celle de la couche avec le plus d'unités dans le module d'association de réponses (*i.e.*, la couche **y***). Ceci est dû au fait que l'on effectue une compression dans le premier module, et que la taille de la couche **y*** est égale à celle de la couche **y**. C'est pourquoi la condition de stabilité devra être testée séparément pour les deux modules.

Une fois de plus, pour s'assurer de minimiser le nombre de postulats techniques du réseau, la même forme de règle sera utilisée pour les deux modules. Ainsi, la règle d'apprentissage pour le module perceptuel devient :

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta_P(\mathbf{y}(0) - \mathbf{y}(t))(\mathbf{x}(0) + \mathbf{x}(t))^{\mathrm{T}} \qquad 5.3$$

$$\mathbf{V}(k+1) = \mathbf{V}(k) + \eta_P(\mathbf{x}(0) - \mathbf{x}(t))(\mathbf{y}(0) + \mathbf{y}(t))^{\mathrm{T}} \qquad 5.4$$

où $\mathbf{W}(k)$ et $\mathbf{V}(k)$ représentent les contenus des matrices de poids de connexion du module perceptuel lors de l'essai d'apprentissage $k$, $\mathbf{x}(0)$ représente l'entrée initiale du réseau, $\mathbf{y}(0)$

[41] Cette propriété spécifique ne sera pas étudiée ici.

représente la compression initiale, $\mathbf{y}(t)$ et $\mathbf{x}(t)$ représentent les vecteurs d'état finaux après $t$ itérations dans le réseau, et $\eta_P$ est un paramètre général d'apprentissage pour ce module. La règle du module d'association de réponses est définie comme suit :

$$\mathbf{Q}(k+1) = \mathbf{Q}(k) + \eta_A(\mathbf{z}(0) - \mathbf{z}(t))(\mathbf{y}^*(0) + \mathbf{y}^*(t))^T \quad 5.5$$

$$\mathbf{P}(k+1) = \mathbf{P}(k) + \eta_A(\mathbf{y}^*(0) - \mathbf{y}^*(t))(\mathbf{z}(0) + \mathbf{z}(t))^T \quad 5.6$$

où $\mathbf{Q}(k)$ et $\mathbf{P}(k)$ représentent les contenus des matrices de poids de connexion du module d'association de réponses lors de l'essai d'apprentissage $k$, $\mathbf{y}^*(0)$ représente la copie initiale de la compression $\mathbf{y}(t)$, $\mathbf{z}(0)$ représente l'entrée initiale du réseau au niveau de la couche de réponse, $\mathbf{y}^*(t)$ et $\mathbf{z}(t)$ représentent les vecteurs d'état finaux après $t$ itérations dans le réseau, et $\eta_A$ est un paramètre général d'apprentissage pour ce module.

### 5.1.4 Tâches possibles pour le modèle

Ce second modèle sera utilisé dans des tâches d'identification et de catégorisation perceptuels avec rétroaction externe. Tel que soutenu par les résultats de Schyns et Rodet (1997), durant l'apprentissage catégoriel, le module perceptuel continuera à effectuer de l'extraction de composantes[42]. La taille de la couche de réponse du second module dépendra du nombre d'alternatives, *i.e.* du nombre d'exemplaires (identification) ou de catégories (catégorisation).

## 5.2 Simulation : Apprentissage et rappel identificatif avec rétroaction externe

Cette première simulation visait à vérifier le caractère adéquat de la procédure d'apprentissage et de rappel utilisant les deux modules. Aussi, elle visait à vérifier que les conditions de départ du réseau (en particulier, le nombre d'unités de compression) étaient appropriées pour la réalisation de la simulation suivante.

[42] Dans le cadre de cette thèse, la possibilité de récupérer les composantes perceptuelles associées à une catégorie (traitement descendant) ne sera cependant pas testée.

### 5.2.1 Méthodologie

Pour cette simulation, des stimuli représentant divers objets en tons de gris (Figure 5.3) ont été utilisés. Cet ensemble de stimuli a été choisi parce que l'on connaît le nombre minimal d'unités de compression nécessaire (32 unités) à leur apprentissage perceptuel parfait (voir Section 4.2.2). Le critère d'arrêt pour l'apprentissage des associations « compression-réponse » est en fait moins strict que le critère d'apprentissage perceptuel, car la couche de réponse **z** contient moins d'unités que la couche de reconstruction **x**. Ce nombre minimal d'unités de compression (32 unités) fut donc utilisé pour la simulation.

Les réponses désirées du réseau furent définies à l'aide de vecteurs de 10 positions. Pour chaque vecteur, l'une des positions avait une valeur égale à 1, alors que toutes les autres valeurs étaient égales à -1. La correspondance entre les stimuli et les vecteurs de réponse est illustrée à la Figure 5.3. Les cases noires correspondent à une valeur de 1, et les cases blanches à une valeur de -1[43].

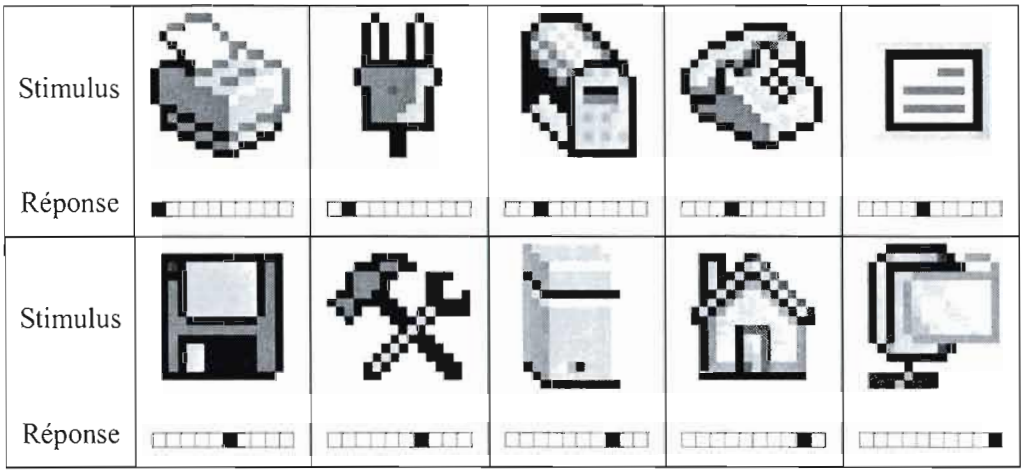


**Figure 5.3** Correspondances entre les entrées et les réponses pour cette simulation.

Les paramètres de FEBAM-RA furent fixés aux valeurs suivantes : le paramètre général de transmission fut fixé à $\delta = 0.1$ et les paramètres général d'apprentissage furent fixés à $\eta_P = 0.001$ (valeur maximale : 0.002) et $\eta_A = 0.005$ (valeur maximale : 0.019). Pour cette

[43] Vu la capacité du BHM de créer des états non-binaires dans le réseau, le choix de la codification exacte des réponses n'est pas crucial pour la performance.

simulation, le critère d'arrêt fut déplacé à la couche **z**. On mesure ainsi la différence entre le vecteur de réponse **z**($t$) et la réponse désirée **z**(0). Ce critère d'adéquation utilise toujours l'erreur quadratique, qui doit ici être calculée comme suit :

$$EQ = \left[ (\mathbf{z}(1) - \mathbf{z}(0))^T (\mathbf{z}(1) - \mathbf{z}(0)) \right] / H \qquad 5.7$$

où $H$ représente le nombre d'unités de la couche de réponse.

La procédure d'apprentissage fut la suivante :

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]), et initialisation des poids à zéro dans les matrices **P** et **Q** (BHM : Chartier et Boukadoum, 2006)[44];
1. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le module perceptuel, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 5.3 et 5.4;
   d. Transmission identitaire (copie) de la compression finale **y**(1) vers la couche **y***.
   e. Réalisation d'un cycle dans le module d'association de réponse, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 5.1 et 5.2), à l'aide du vecteur de réponse correspondant;
   f. Mise à jour des matrices de poids de connexion **P** et **Q** selon les équations 5.5 et 5.6;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Calcul de l'erreur quadratique moyenne (au niveau de la couche de réponse) pour l'ensemble de stimuli;
4. Répétition des étapes 1 à 3 jusqu'à ce que l'erreur quadratique moyenne pour un bloc soit inférieure ou égale à $1 \times 10^{-6}$.

La procédure de rappel se déroula comme suit :

1. Essai de rappel :
   a. Sélection d'un stimulus de l'ensemble d'entrée;
   b. Itération dans le module perceptuel (à l'aide des matrices **W** et **V** finales) jusqu'à l'atteinte d'une compression stable (en l'occurrence, lorsque **y**($t$+1) = **y**($t$));
   c. Copie de la compression stable **y**($t$) dans la couche **y***.

[44] Quoiqu'avec le BHM, l'initialisation à zéro ne soit pas obligatoire, dans ce cas-ci, aucun biais de réponse ne sera introduit dans le module préalablement.

d. Itération dans le module d'association de réponse (à l'aide des matrices **P** et **Q** finales) jusqu'à l'atteinte d'une réponse stable (en l'occurrence, lorsque $\mathbf{z}(t+1) = \mathbf{z}(t)$)

2. Répétition de l'étape 1 pour chaque stimulus de l'ensemble.

Lors du rappel, une réponse finale (stable) fut considérée comme acceptable si l'erreur quadratique entre le vecteur de réponse désiré et la réponse du réseau $\mathbf{z}(t)$ était inférieure à $1x10^{-6}$. Un rappel fut déclaré « parfait » lorsque le rappel des 10 vecteurs de réponse originaux respectait cette règle. Pour obtenir une estimation moyenne robuste du nombre de blocs d'apprentissage nécessaire à l'atteinte du critère, la simulation fut répétée 100 fois, en utilisant des poids de départ aléatoires différents pour les matrices du module perceptuel.

### 5.2.2 Résultats

À l'aide de 32 unités de reconstruction, le réseau fut en mesure de réussir le rappel parfait 100% du temps. En moyenne, le nombre de blocs d'essais d'apprentissage nécessaires à l'atteinte du critère fut de 112.89 blocs (avec une erreur-type de 1.7 blocs). Ce résultat contraste avec celui obtenu pour la simulation de reconstruction autonome de ces mêmes stimuli (Figure 4.11). En effet, lorsque l'on utilisait 32 unités de compression, 258.24 blocs d'apprentissage étaient nécessaires, en moyenne, pour la reconstruction parfaite. Cette différence notable est liée au changement de critère durant l'apprentissage; ce dernier est moins strict vu le nombre moindre d'unités dans la couche de réponse **z**. Ceci confirme qu'au niveau du modèle, la différentiation parfaite n'est pas obligatoire pour l'identification (autonome ou supervisée). Le nombre élevé de blocs d'apprentissage requis pour cette tâche permettra possiblement de produire une économie de blocs, suite à une période de pré-exposition aux stimuli. Un nombre de blocs d'apprentissage trop bas ne permettrait pas d'étudier cette capacité du réseau.

## 5.3 Étude : Effet d'une période de pré-exposition perceptuelle

Dans cette étude, le but sera de tester si le modèle FEBAM-RA est en mesure de reproduire (qualitativement) l'effet dit de « Gibson-Walk ». Gibson et Walk (1956) ont testé, à l'aide de rats, l'effet d'une exposition préalable aux stimuli sur la performance de catégorisation. Pour ce faire, ils ont séparé les rats en deux groupes : pour le premier groupe

de rats, les formes utilisées lors de la tâche principale étaient visibles de la cage, durant une période prédéterminée précédant la tâche supervisée. Le deuxième groupe de rats n'était pas exposé à ces images. La tâche supervisée impliquait d'appuyer sur la bonne pédale (et d'ainsi recevoir une récompense) suite à la présentation d'une image (tâche de catégorisation). Les rats ayant été pré-exposés aux stimuli montrèrent une meilleure performance et un apprentissage plus rapide.

D'autres auteurs (dont Gibson, Walk et leurs collègues) ont répliqué cet effet, dans des conditions expérimentales variées. Plusieurs auteurs ont montré que cet effet existait aussi pour les humains. Par exemple, Goss (1953, dans Hall, 1991) a exposé ses participants à différentes intensités lumineuses, préalablement à une tâche de classification de ces intensités. La pré-exposition améliora la performance à la tâche principale. Aussi, Willis et McLaren (1998) ont montré que la pré-exposition au matériel expérimental (des catégories artificielles) pouvait faciliter la catégorisation et la discrimination dans une tâche subséquente. Dans leur cas, la première phase de traitement des stimuli impliquait des jugements de nouveauté, une tâche perceptuelle non-reliée.

Hall (1991; Goldstone, 1998), suite à une recension des multiples réplications expérimentales, affirme que ces dernières ne présentent pas toutes un effet aussi marqué que pour les expériences originales de Gibson et collègues. Il indique toutefois qu'un effet aussi souvent retrouvé doit être considéré comme valide, et que donc, la pré-exposition accélère l'apprentissage dans des tâches d'identification et de catégorisation supervisées. Schyns, Goldstone et Thibaut (1998) rapportent cet effet dans une liste de phénomènes dont un modèle perceptivo-cognitif devrait rendre compte, et rappellent que la rétroaction corrective n'est pas nécessaire pour obtenir des effets de pré-différentiation suite à une exposition.

Dans cette étude, une simple pré-exposition au module perceptuel sera effectuée préalablement à la tâche d'identification supervisée. Ici, l'avantage d'utiliser deux modules distincts prendra tout son sens. Cette division permettra d'entraîner le modèle de façon strictement perceptuelle durant un certain nombre de blocs, sans que ce dernier ne connaisse les associations désirées entre les stimulations perceptuelles et les réponses. Ensuite, utilisant les matrices de poids de connexions (**W** et **V**) développées durant la pré-exposition, on pourra poursuivre l'apprentissage de façon supervisée, en utilisant les deux modules. Durant cette

seconde phase, puisque la réponse sera fournie de façon externe, le modèle pourra associer des vecteurs-réponses aux représentations comprimées, dont le développement sera en stade avancé.

### 5.3.1 Méthodologie

Les stimuli (images en tons de gris), les vecteurs de réponse, le nombre d'unités de compression, et les paramètres du modèle furent identiques à ceux utilisés pour la simulation précédente. La procédure d'apprentissage se déroula comme suit :

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]), et initialisation des poids à zéro dans les matrices **P** et **Q**;
   a. Si une phase de pré-exposition doit avoir lieu, passer à l'étape 1; sinon, passer à l'étape 4.
1. Essai de pré-exposition :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le module perceptuel, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 5.3 et 5.4;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Répétition des étapes 1 et 2 pour un nombre prédéterminé de blocs (selon la condition : 20, 40 ou 60 blocs).
4. Essai d'apprentissage :
   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le module perceptuel, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 5.3 et 5.4;
   d. Transmission identitaire de la compression finale $\mathbf{y}(1)$ dans la couche $\mathbf{y}^*$.
   e. Réalisation d'un cycle dans le module d'association de réponse, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 5.1 et 5.2), à l'aide du vecteur de réponse correspondant;
   f. Mise à jour des matrices de poids de connexion **P** et **Q** selon les équations 5.5 et 5.6;
5. Répétition de l'étape 4 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
6. Calcul de l'erreur quadratique moyenne (au niveau de la couche de réponse) pour l'ensemble de stimuli;

7. Répétition des étapes 4 à 6 jusqu'à ce que l'erreur quadratique moyenne pour un bloc soit inférieure ou égale à $1 \times 10^{-6}$.

Pour obtenir des estimations robustes du nombre de blocs nécessaires à l'atteinte du critère de réponse, 100 groupes de quatre simulations furent effectués. Un groupe de quatre simulations couvrait chacune des quatre conditions (0, 20, 40 ou 60 blocs de pré-exposition). Les quatre simulations étaient liées en ce qu'elles utilisaient les mêmes matrices de poids de départ au niveau perceptuel (**W** et **V**). Aussi, la séquence aléatoire de présentation des exemplaires fut la même pour les quatre simulations d'un groupe. La différence entre ces simulations repose sur le stade de développement des matrices **W** et **V** lorsque la période de pré-exposition prit fin. En effet, puisque la mise à jour de ces matrices est effectuée durant la pré-exposition (tout comme dans FEBAM : apprentissage autonome), ces matrices différaient lorsque l'apprentissage subséquent débuta.

### 5.3.2 Résultats

Des résultats représentatifs obtenus pour l'apprentissage normal (deux modules) sont présentés à la Figure 5.4. Ces résultats sont tirés d'une seule simulation, mais présentent des différences qualitatives qui ont été observées pour toutes les réplications réalisées. Ici, sur chaque graphe, on compare la courbe d'apprentissage pour la condition sans pré-exposition, avec une courbe obtenue dans l'une des conditions avec pré-exposition. On peut ici observer que la pré-exposition réduit le nombre de blocs nécessaire à l'apprentissage des réponses, et ce, dans toutes les conditions. Aussi, la courbure de la courbe d'apprentissage pour le module d'association de réponse devient plus prononcée suite à une pré-exposition plus longue; ceci dénote un apprentissage plus rapide dès les premiers stades d'entraînement. Cette différence est particulièrement apparente lorsque l'on compare les courbes suivant 20 blocs (Figure 5.4(a), ligne pleine) et 40 blocs (Figure 5.4(b), ligne pleine) de pré-exposition.

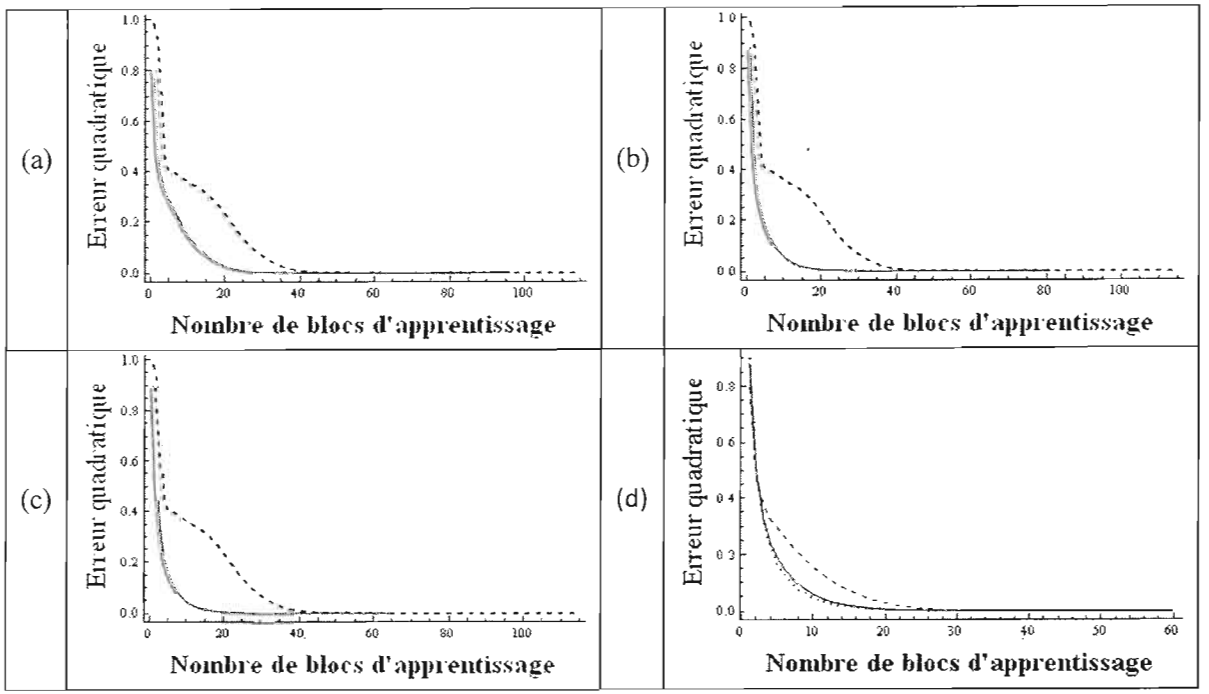


**Figure 5.4** Courbes d'apprentissage représentatives (une seule simulation) pour l'apprentissage utilisant les deux modules. Sur les trois graphes, la courbe en pointillé représente l'apprentissage du module d'association de réponse pour la condition sans pré-exposition. La courbe pleine représente l'apprentissage du module d'association de réponse lorsque l'on pré-entraîne le module perceptuel pour (a) 20 blocs; (b) 40 blocs; (c) 60 blocs. On voit ici que la pré-exposition réduit le nombre de blocs nécessaire pour atteindre le critère. (d) Courbes d'apprentissage comparatives pour 20 blocs (ligne pointillée), 40 blocs (ligne pleine), et 60 blocs (ligne en petits points). Seuls les 60 premiers blocs d'apprentissage normal sont présentés.

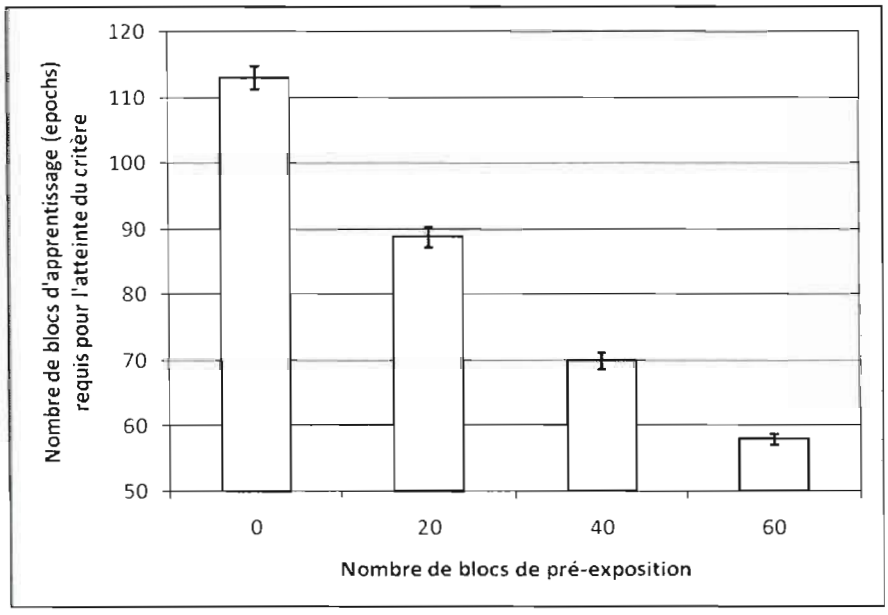


**Figure 5.5** Nombre de blocs moyen requis pour atteindre le critère d'apprentissage, en fonction du nombre de blocs de pré-exposition effectués à l'aide du module perceptuel. Les barres d'erreur représentent une erreur-type.

Les résultats globaux pour la simulation sont présentés à la Figure 5.5. Le nombre de blocs d'apprentissage requis pour l'atteinte du critère d'apprentissage furent, en moyenne, de 113.06 blocs pour la condition sans pré-exposition, de 88.79 blocs pour la condition avec 20 blocs de pré-exposition, 69.92 blocs pour la condition avec 40 blocs, et 57.96 blocs pour la condition avec 60 blocs. La valeur des erreurs-type est comprise entre 0.80 (60 blocs) et 1.76 (aucune pré-exposition). En observant les différences, on voit ici que la simple pré-exposition au module perceptuel produit un net effet sur la rapidité de l'apprentissage identificatif.

On peut aussi observer que l'effet du pré-entraînement perceptuel s'amenuise avec le nombre de blocs de pré-exposition. La différence entre une absence de pré-exposition et une pré-exposition de 20 blocs est de 24.27 blocs d'apprentissage. En comparaison, la différence entre les conditions 40 blocs et 60 blocs n'est que de 11.96 blocs d'apprentissage. Cet amenuisement est dû à l'utilité décroissante de la pré-exposition. En effet, lors de l'apprentissage perceptuel, l'erreur quadratique (en fonction du nombre de blocs

d'apprentissage) suit une courbe de puissance. Ceci implique que l'amélioration de la performance du module perceptuel est beaucoup plus marquée durant les premiers blocs d'apprentissage. Le module d'association de réponse peut donc rapidement se fier sur des représentations comprimées de meilleure qualité. Une période de pré-exposition supplémentaire améliore la qualité de ces compressions, mais de moins en moins à mesure que le nombre de blocs d'apprentissage augmente.

Cette simulation montre donc la capacité du modèle FEBAM-RA à reproduire l'effet Gibson-Walk. Aussi, cela valide l'utilisation de deux modules qui effectuent séparément le traitement au niveau perceptuel, et l'association des compressions avec des vecteurs de réponse prédéterminés. Une architecture à deux modules, comme celle de FEBAM-RA, permet donc de reproduire le processus de la tâche de façon fidèle.

## 5.4 Étude: Comparaison des processus d'identification/catégorisation supervisés

Dans une série d'expériences utilisant deux tâches simultanées (*dual-task*), Reed (1978) a comparé la performance à des tâches d'identification et de catégorisation perceptuelle avec rétroaction corrective. Il a demandé à des participants d'apprendre les étiquettes identificatives et catégorielles reliées à deux catégories d'exemplaires, créées autour de prototypes. Lors de chaque essai, un participant, suite à la présentation d'un exemplaire, devait répondre en fournissant le numéro lié à l'exemplaire, et la lettre liée à la catégorie. Reed a montré que de façon longitudinale, l'appartenance catégorielle était toujours apprise plus rapidement que l'étiquette catégorielle. Selon Reed, ce résultat pouvait être utilisé pour invalider l'approche exemplariste, puisque l'apprentissage plus lent des exemplaires exclut leur récupération comme processus de base de la catégorisation.

Cette dernière simulation vise deux buts. Premièrement, on voudra ici montrer, une fois de plus, que la distinction entre les processus au niveau objet (tels que l'identification) et au niveau catégorie peuvent en fait n'être que les deux extrêmes d'un continuum quantitatif. Seul le niveau de distribution de la représentation comprimée fait une différence. Deuxièmement, on voudra s'assurer que dans la mesure où l'on voudra éventuellement modéliser la dissociation entre les systèmes cognitifs « objet » et « catégorie » (tel que

proposée par Knowlton et Squire, 1993), un modèle basé sur cette distinction pourra respecter un fait établi : les associations catégorielles sont apprises plus rapidement que les associations identificatives, lorsque les catégories sont construites autour de prototypes (Reed, 1978).

### 5.4.1 Méthodologie

Deux catégories, bâties autour de prototypes, furent créées pour cette simulation. Chaque vecteur prototypique, d'une taille de 100 positions, fut généré aléatoirement, avec la contrainte que la corrélation moyenne entre les prototypes devait être égale à $r = 0.36$. Pour chacune des catégories, cinq exemplaires furent créés en inversant la valeur de 16 pixels à partir du prototype. Chaque ensemble de pixels inversés était exclusif. Les corrélations intra-catégorielles (entre exemplaires d'une même catégorie) moyennes étaient égales à $r = 0.24$, tout comme corrélations inter-catégorielles (entre exemplaires de catégories opposées) moyennes. Le fait que ces deux dernières corrélations soient identiques rendent l'apprentissage catégoriel beaucoup plus difficile, et permettent d'utiliser des stimuli dont la similarité respecte l'esprit des ensembles d'images utilisés par Reed (1978). Ce dernier a utilisé des visages de Brunswik (Reed, 1972) pour ses expériences. Pour ces stimuli, il était impossible, sans apprentissage supervisé, de déterminer une appartenance catégorielle claire pour chaque exemplaire. Il est à noter que seuls les 10 exemplaires furent présentés à l'apprentissage et au rappel; les prototypes d'origine ne furent jamais présentés. Les deux prototypes et les dix exemplaires utilisés sont illustrés à la Figure 5.6.

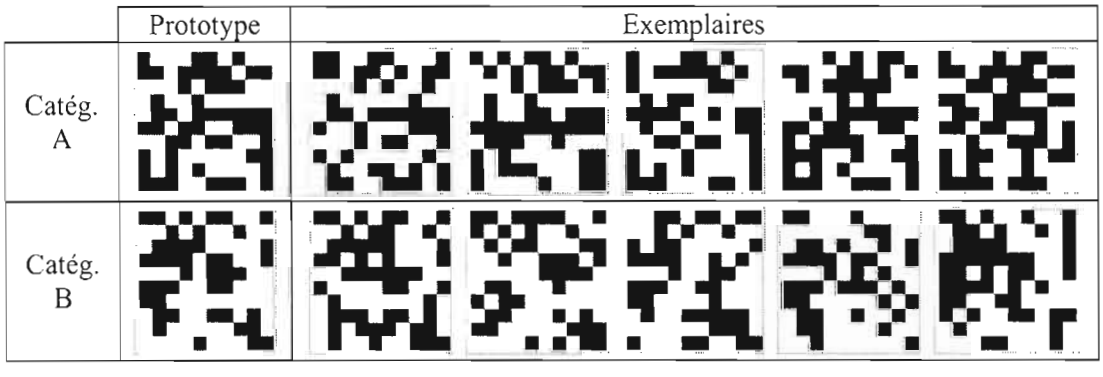


**Figure 5.6** Prototype générés aléatoirement, et exemplaires dérivés, pour chacune des deux catégories.

Par fidélité à la théorie des systèmes multiples, pour chacune des simulations, deux sous-systèmes furent postulés, et utilisés simultanément. Chaque sous-système est en fait un modèle FEBAM-RA. Le premier de ceux-ci, le sous-système « objet », servait à associer des compressions avec des vecteurs de réponse identificatifs (un vecteur de réponse par stimulus). Le second sous-système « catégorie » servait à associer ces mêmes compressions avec des vecteurs de réponse catégoriels (un vecteur de réponse distinct par catégorie). Tout comme pour la simulation de catégorisation autonome, un procédé d'ajouts d'unités de compression fut utilisé. Chaque réseau, au départ, ne présentait qu'une unité dans les couches **y** et **y***. Les correspondances entre les stimuli et les vecteurs de réponse identificatifs et catégoriels sont illustrées à la Figure 5.7.

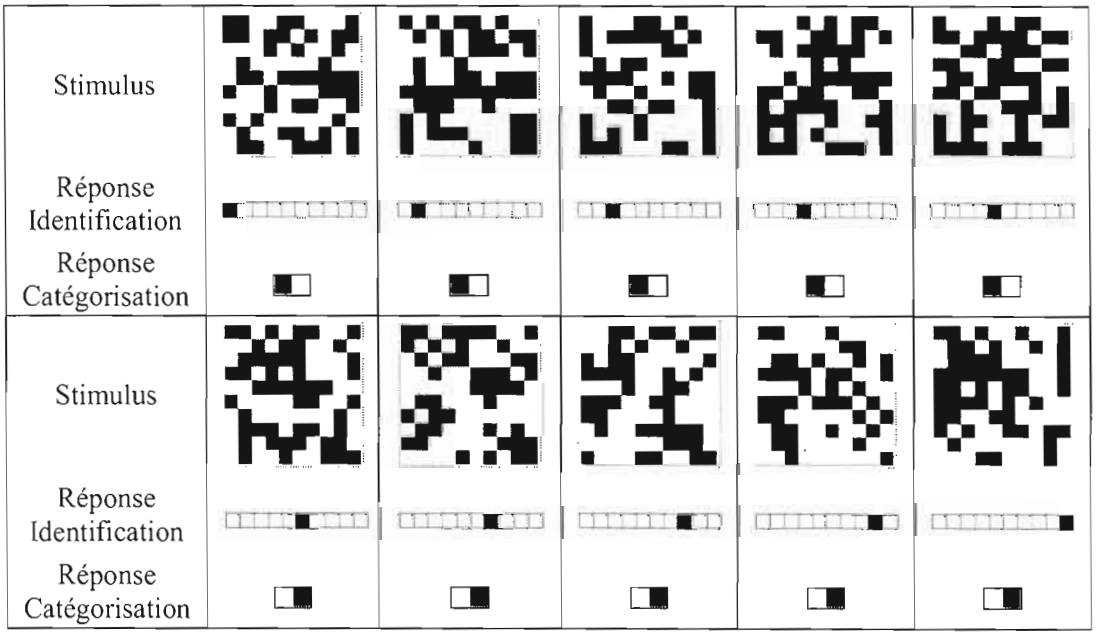


**Figure 5.7** Correspondances entre les entrées et les réponses pour cette simulation.

La procédure de rappel fut la même qu'à la section 5.2.1. La procédure d'apprentissage utilisée pour chaque sous-système fut la suivante :

0. Initialisation aléatoire des poids de connexion dans les matrices **W** et **V** (Intervalle des valeurs de départ : [-0.1, 0.1]), et initialisation des poids à zéro dans les matrices **P** et **Q**;
1. Essai d'apprentissage :

   a. Sélection aléatoire d'un vecteur d'entrée parmi l'ensemble de départ;
   b. Réalisation d'un cycle dans le module perceptuel, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 3.1 et 3.2);
   c. Mise à jour des matrices de poids de connexion **W** et **V** selon les équations 5.3 et 5.4;
   d. Transmission identitaire de la compression finale **y**(1) dans la couche **y***.
   e. Réalisation d'un cycle dans le module d'association de réponse, tel qu'illustré à la Figure 5.2 (utilisant les fonctions de transmission décrites aux équations 5.1 et 5.2), à l'aide du vecteur de réponse correspondant;
   f. Mise à jour des matrices de poids de connexion **P** et **Q** selon les équations 5.5 et 5.6;
2. Répétition de l'étape 1 jusqu'à ce que chaque stimulus de l'ensemble ait été traité par le réseau (apprentissage par blocs ou *epochs*);
3. Calcul de l'erreur quadratique moyenne (au niveau de la couche de réponse) pour l'ensemble de stimuli;
4. Répétition des étapes 1 à 3 jusqu'à ce que la différence d'erreur quadratique moyenne entre deux blocs consécutifs d'apprentissage soit inférieure ou égale à $1x10^{-6}$.
5. Ajout d'une unité dans les couches **y et y***;
6. Initialisation aléatoire des poids de connexion pour la nouvelle colonne ajoutée à la matrice **W** et la nouvelle ligne ajoutée à la matrice **V** (Intervalle des valeurs de départ : [-0.1, 0.1]), et initialisation des poids à zéro pour la nouvelle colonne ajoutée à la matrice **P** et la nouvelle ligne ajoutée à la matrice **Q**;
7. Répétition des étapes 1 à 6 jusqu'à ce que l'erreur quadratique moyenne soit inférieure ou égale à $1x10^{-6}$.

La « liaison » entre les sous-systèmes d'une même simulation implique que ces deux derniers utilisaient les mêmes poids de connexion au départ, et que l'ordre de présentation aléatoire des stimuli fut le même pour les deux sous-systèmes. Ceci respecte l'esprit des tâches de Reed, où on demandait aux participants de fournir deux décisions (numéro d'exemplaire et lettre associée à la catégorie) lors de chaque essai. Aussi, par souci de contrôle, les deux sous-systèmes utilisèrent les mêmes paramètres. Dans les deux sous-systèmes, le paramètre $\eta_P$ était égal à 0.00625, et le paramètre $\eta_A$ était égal à 0.0625. Le paramètre $\delta$ était, comme toujours, égal à 0.01. Le choix de paramètres identiques pour les deux sous-systèmes permet de limiter l'influence de ceux-ci sur les comparaisons à effectuer. Ainsi, la différence de performance entre les sous-systèmes ne devrait être due qu'à la

difficulté des différentes tâches[45]. 100 simulations furent réalisées. Pour chaque simulation, les poids de départ et l'ordre de présentation des stimuli étaient différents.

### 5.4.2 Résultats

Les sous-systèmes furent en mesure de réussir le rappel parfait 100% du temps. Du point de vue de la phase d'apprentissage, deux variables peuvent être comparées entre les sous-systèmes : 1) le nombre de blocs d'apprentissage nécessaire à l'atteinte du critère, et 2) le nombre d'unités de compression nécessaire pour pouvoir apprendre correctement les associations entre les compressions et les vecteurs-réponses. Dans les deux cas, les moyennes sont nettement plus élevées pour la tâche d'identification. En moyenne, le sous-système « objet » nécessite 321 blocs d'apprentissage pour atteindre le critère (erreur-type : 9.28 blocs), alors que le sous-système « catégorie » n'a besoin que de 209.9 blocs (erreur-type : 7.68 blocs). Ceci tombe sous le sens, puisque le nombre d'alternatives à départager est moindre dans le cas de la tâche de catégorisation. En ce qui a trait au nombre d'unités de compression, on retrouve une aussi différence marquée en faveur du sous-système « catégorie », qui nécessite en moyenne 12.87 unités de compression (erreur-type : 0.63 unités), alors que le sous-système « objet » en utilise en moyenne 19.01 (erreur-type : 0.34), Ce résultat est également parfaitement intuitif : puisque le nombre de séparations nécessaires dans l'espace-réseau est plus élevé pour la tâche d'identification, le nombre de dimensions de l'espace (unités de compression) doit également être plus élevé pour permettre la possibilité dc créer des hyperplans de séparation adéquats.

L'effet montré par Reed (1978) est ici reproduit, puisque toutes les variables étudiées portent à croire que la tâche de catégorisation est plus facile pour le réseau que la tâche d'identification. Aussi, une fois de plus, cette simulation montre qu'il est possible de modéliser des tâches perceptuelles au niveau « objet » et au niveau « catégorie » à l'aide d'un seul modèle, pour lequel seul le nombre d'unités de compression diffère. Ceci constitue une

[45] Ceci n'implique pas que l'auteur supporte une position théorique où ces deux systèmes, au niveau cognitif, apprendraient à la même vitesse. Il est clair que dû aux limites du système de traitement humain (Goldstein, 2008), l'accent mis sur l'utilisation de l'un des sous-systèmes devrait être fait au détriment de la rapidité d'apprentissage du second.

simple différence quantitative entre les deux tâches. Les principes de base du modèle et les conditions de départ utilisées sont identiques.

## 5.5 Conclusion

Le modèle FEBAM-RA étend les possibilités d'identification et de catégorisation de FEBAM pour le cas où une rétroaction externe est disponible. Cette extension se fait sans aucun postulat supplémentaire, que ce soit au niveau de l'architecture du module, du type d'apprentissage, de la règle de transmission, et de la règle d'apprentissage. Ainsi, l'approche FEBAM-RA est nettement plus parcimonieuse que n'importe quelle approche dite « hybride », basée sur la mise en commun de principes de base et de postulats provenant de différentes classes de réseaux.

Le modèle, dans sa forme actuelle, est en mesure de reproduire l'effet Gibson-Walk, qui implique une amélioration de la performance suite à une exposition préalable aux stimuli de la tâche. Cette amélioration est plus marquée pour les premiers blocs de pré-exposition; ceci est dû à la réduction de l'erreur quadratique au niveau perceptuel, qui suit une courbe de puissance. Tel qu'argumenté, la réplication de cet effet est rendue possible en postulant deux modules séparés : un module perceptuel autonome, et un module d'association de réponse. Ce second module n'est sollicité que lorsqu'une rétroaction externe est disponible.

Finalement, la dernière simulation confirme que l'identification et la catégorisation, du point de vue du modèle, peuvent être vues comme des processus qualitativement identiques, différant simplement au niveau de la distribution des représentations. Les résultats de cette simulation montrent une réplication de l'effet de difficulté lors de réalisation de tâches perceptuelles simultanées. Ces résultats permettent aussi de croire que l'utilisation de deux sous-systèmes, en accord avec la théorie des systèmes multiples (niveau objet vs. niveau catégorie), pourrait mener à la modélisation adéquate de résultats expérimentaux nécessitant les deux types de traitement perceptuel.

Il convient de noter que la proposition présentée dans ce chapitre constitue un essai exploratoire, et que les résultats possèdent donc un caractère provisoire. En effet, l'étude du « processus inverse » de catégorisation allant de la réponse vers la motricité est peu

développée. Cependant, il serait intéressant de continuer la réflexion théorique et l'étude empirique du phénomène, ainsi que l'étude du processus liant les catégories finales à la récupération des caractéristiques définitoires.

# CHAPITRE VI

# DISCUSSION GÉNÉRALE

## 6.1 Contributions à souligner

Plusieurs contributions intéressantes émanent de ce projet de recherche doctoral. Pour débuter, on doit souligner le fait que l'on ait ici remis l'accent sur l'importance du processus de formation autonome d'un vocabulaire de composantes. Ce processus de traitement ascendant a été proposé par Harnad (1990) et Goldstone, Schyns et Thibaut (1998), et a été validé entre autres par les études de Schyns et collègues (Schyns et Murphy, 1991, 1994; Schyns et Rodet, 1997). Même si l'on devrait toujours en tenir compte lorsque l'on propose un modèle d'apprentissage ou de catégorisation perceptuel, la plupart des modèles (principalement les modèles symboliques) en font abstraction. Bien que les mémoires de composantes soient pour l'instant peu interprétables dans FEBAM, elles sont facilement récupérables, sans analyse mathématique supplémentaire : ce sont les poids de connexion de la matrice de reconstruction. Des simulations sont présentement en cours, pour déterminer si l'utilisation de stimuli basés sur un vocabulaire de composantes orthogonales (Schyns et Rodet, 1997) permettent plus facilement de récupérer des composantes visuellement significatives.

Les autres contributions qui seront énumérées ici sont liées à l'effort de simplicité et de parcimonie qui transcende l'approche proposée. De nombreux processus et effets ont été modélisés, et ce, à l'aide d'un minimum de postulats techniques et théoriques. Pour débuter, le modèle FEBAM réussit, en une seule opération, à effectuer la réduction dimensionnelle (menant à une nette économie cognitive) et le développement de caractéristiques. Ceci est rendu possible grâce à l'utilisation conjointe de deux matrices de poids de connexion, l'une

servant à la compression, et l'autre, à la reconstruction. Ces deux matrices sont mises à jour en utilisant une seule règle d'apprentissage.

Utilisant ces mêmes principes définitoires, on peut aussi modéliser deux tâches supplémentaires, soit la catégorisation et l'identification autonome. Ici, si le réseau considère que deux stimulations perceptuelles menant à la même compression au rappel sont indistinctes. C'est soit qu'il s'agit du même objet (identification), ou que ces stimulations font partie de la même catégorie. Ces deux processus liés au niveau objet et au niveau catégorie peuvent être modélisés en modifiant une seule variable, soit la distribution des représentations (nombre d'unités de compression). Ceci simplifie la façon de voir les processus « objet » et « catégorie », qui, plutôt que d'être vus comme qualitativement différents, ne diffèrent qu'au niveau quantitatif, sur un continuum de distribution des représentations. On peut donc modéliser les deux types de tâches sans aucun ajout de postulat.

Également, en utilisant le lien qui unit les RAM et les BAM (la RAM .étant une spécification de la BAM), il a été possible d'étendre le modèle au cas supervisé en ajoutant un second module. Ce module utilise la même architecture, la même règle de transmission, et la même règle d'apprentissage que FEBAM. La seule différence entre les deux modules est le nombre de points d'entrée du réseau. Le module d'association de réponse étant en fait une BHM, celui-ci montre deux points d'entrée, l'un du côté de la compression et l'autre du côté de la réponse. On a aussi montré que pour les tâches supervisées, la différence entre le niveau objet et le niveau catégorie n'était une fois de plus lié au nombre d'unités de compression.

## 6.2 Perspectives

Dans sa forme actuelle, le réseau FEBAM réussit à reproduire, de façon qualitative, plusieurs tâches et effets perceptivo-cognitifs retrouvés dans la littérature. Cela dit, le but de la thèse était de débuter de la façon la plus simple possible. On voulait ici étudier un maximum de possibilités et de caractéristiques sans ajouter un nombre incalculable de postulats. L'approche FEBAM devra bien évidemment être ajustée et complexifiée, si l'on veut à moyen terme rendre compte de données empiriques, et pouvoir présenter des mesures

d'adéquation quantitatives. Voici donc, pour conclure, quelques propositions de futurs travaux permettant d'améliorer le modèle.

### 6.2.1 Apprentissage bruité

Au Chapitre 4, on a soutenu que le rappel d'entrées bruitées ne pouvait faire de sens au niveau perceptivo-cognitif que dans la mesure où ces mêmes entrées bruitées étaient présentées à l'apprentissage. En effectuant ceci, on postule que le système doit tenir compte d'un certain niveau de bruit interne, ainsi que d'une variance perceptuelle liée aux entrées.

Ce principe de variance perceptuelle est supporté par Ashby et Lee (1993), qui considèrent qu'un stimulus ne sera jamais perçu exactement de la même façon d'une présentation à l'autre, dû à des conditions externes, mais aussi internes au système. Ainsi, selon eux, la variabilité, cette « vérité fondamentale » de la perception, doit être prise en compte par tout modèle perceptuel. Dans FEBAM, la variabilité peut être modélisée grâce à l'ajout de bruit gaussien à l'entrée. L'utilisation de ce type de bruit peut être justifié de plusieurs façons, l'une d'elles étant que l'échantillonnage de plusieurs types de bruits internes et externes lors du traitement devrait, selon le théorème central limite, mener à une distribution normale du bruit traité.

L'une des caractéristiques intrinsèques de FEBAM est la possibilité d'apprendre à partir d'entrées bruitées. Cette caractéristique n'a pas été explorée dans la thèse, mais des essais ont été réalisés par Giguère, Chartier, Proulx et Lina (2007a). Ils ont effectué une phase d'apprentissage avec les stimuli de la section 4.3 (alphabet de 26 lettres), et ont montré qu'en ajoutant 20% de bruit aux entrées à l'apprentissage et au rappel, le modèle était en mesure de parfaitement reconstruire les stimuli originaux non-bruités. Des comparaisons avec des modèles à base de PCA, nPCA et ICA ont montré que ces modèles ne suffisaient pas à la tâche.

Évidemment, tel que déjà mentionné, le but du système n'est pas la reconstruction parfaite, mais la différentiation au niveau objet et catégorie. De nouvelles simulations seront donc nécessaires pour explorer le comportement du réseau en ce qui a trait à la catégorisation et l'identification autonome ou supervisée à partir d'entrées bruitées.

### 6.2.2 Temps de traitement « réaliste »

Avec la grande majorité des modèles de type RAM ou BAM, on présente une seule image (patron d'entrée), et cette image effectue un certain nombre d'itérations dans la ou les matrices de poids de connexions avant leur mise à jour. Ce principe répandu ne correspond peut-être pas à la réalité du système perceptivo-cognitif. Ce système est nourri d'images de façon continue, et nous possédons des évidences empiriques permettant de croire que seules quelques-unes de ces entrées peuvent être traitées durant chaque seconde d'exposition. Ainsi, chaque objet peut être vu comme une séquence d'images de ce même objet.

Pour déterminer de façon approximative le nombre d'images pouvant être traitées par un système perceptivo-cognitif, on doit explorer des travaux provenant de deux domaines, soit ceux de la mémoire et de l'attention. Côté mémoire, nous savons que les humains possèdent une mémoire iconique, contenant une représentation non-interprétée du dernier champ visuel enregistré. Cette mémoire se nomme le registre d'information sensoriel (Sperling, 1960; Goldstein, 2008). Au niveau visuel, « l'impression » d'une image dans ce système dure approximativement 200-250 millisecondes.

L'existence du registre d'information sensoriel supporte le principe de comparaison entre l'entrée et la reconstruction dans FEBAM. Pour effectuer cette comparaison, on doit postuler que l'entrée originale est toujours disponible au système. Si l'on se fie aux travaux de Sperling, cette entrée est effectivement disponible, mais pour une durée limitée. Cela signifie que le processus théorique effectué par FEBAM peut durer au maximum un quart de seconde. Ainsi, si l'on veut traduire en temps expérimental réel, en moyenne, quatre à cinq images pourraient être traitées par le système durant cette période (approx. 200-250 millisecondes chacune).

Les recherches en attention supportent aussi cette fréquence de traitement. Raymond, Shapiro et Arnell (1992) ont proposé le principe du vacillement attentionnel (*attentional blink*). Selon ce principe, lorsque le système perceptivo-cognitif porte attention aux entrées perceptuelles visuelles (ce qui, dans FEBAM, est pris pour acquis), il semble y avoir une période réfractaire durant laquelle les entrées ne sont pas traitées. Cette période dure entre 200 et 500 millisecondes. La période la plus affectée par le vacillement attentionnel tourne

également autour de 225-250 millisecondes. Durant cette période, il est généralement reconnu qu'un traitement perceptuel est en cours, et que le système de traitement « ferme la porte » à d'autres entrées pour éviter une certaine confusion perceptuelle. Du côté de FEBAM, ce « traitement » perceptuel pourrait être la mise à jour des poids de connexion.

Ainsi, on pourrait à l'avenir tenter de modéliser chaque essai de traitement perceptuel comme une séquence de versions bruitées provenant du même stimulus. Ceci respecterait le principe de variance perceptuelle, et si l'on postule que le système traite 4 à 5 de ces entrées par seconde, pourrait constituer une avancée vers une modélisation plus réaliste au niveau des temps de traitement.

### 6.2.3 Procédure de vigilance

Aux sections 4.3.2, 4.3.3, et 5.4, on a utilisé un processus itératif d'ajout d'unités de compression dans le réseau. Ce procédé permettait d'augmenter le nombre de représentations possibles distinctes des objets et des catégories dans le réseau. La décision d'ajouter une unité était basée sur une stabilisation de l'erreur : dans le cas autonome, on utilisait l'erreur de compression, et dans le cas supervisé, l'erreur de réponse. Évidemment, ces critères, bien qu'adéquats pour l'exploration technique des capacités de l'approche, devront être modifiés à court ou moyen terme.

Une approche bien connue qui nous permettrait de déterminer quand le modèle autonome doit recruter une unité a été proposée par Grossberg et collègues pour la famille de réseaux ART. Les réseaux ART utilisent un paramètre de vigilance pour déterminer la nécessité d'associer un stimulus donné à une unité de sortie (représentation localiste). Lors de chaque essai, on vérifie la corrélation entre le vecteur d'entrée, et chaque vecteur de sortie disponible. Si la corrélation est plus basse que le paramètre de vigilance, alors le réseau recrute une unité de sortie, et cette unité représentera le vecteur d'entrée. Dans ce cas, on postule des représentations localistes, ce qui va à l'encontre des postulats de FEBAM.

Hélie, Chartier et Proulx (2006) ont proposé un principe similaire s'appliquant aux modèles RAM, à représentations distribuées. Lorsqu'un stimulus est présenté au réseau, il itère jusqu'à atteinte d'un point stable. Si la corrélation entre le vecteur-stimulus et

l'attracteur stable atteint dépasse un certain critère, alors l'attracteur reste inchangé. Si la corrélation est trop basse, alors la position de l'attracteur est modifiée; on effectue une moyenne pondérée entre le vecteur d'entrée et le vecteur représentant l'attracteur.

Pour FEBAM, le paramètre de vigilance ne serait pas utilisé pour déplacer des positions d'attracteurs, ou pour fixer directement une appartenance identificative ou catégorielle, mais plutôt pour déterminer la nécessité d'ajouter des unités de compression. Tel que montré au Chapitre 4, lorsque les catégories sont relativement bien séparées, l'ajout d'unités dans le réseau permet d'augmenter le nombre de divisions catégorielles (jusqu'à l'atteinte d'une différentiation au niveau « objet »), mais le paysage catégoriel demeure relativement stable d'un bloc au suivant. Ainsi, le simple recrutement d'unités serait suffisant pour permettre au système de préciser le paysage catégoriel, sans nécessairement perdre les régularités statistiques déjà extraites.

Ce procédé est lié à plusieurs avantages : pour débuter, nous savons que les participants aux expériences de catégorisation ne montrent pas des patrons de réponse constants, qui nous permettraient de croire que l'appartenance catégorielle est fixée aussi rapidement que pour les réseaux ART. Même si l'appartenance catégorielle d'un item est adéquate, rien ne dit que le participant réussira à classifier cet item correctement au prochain bloc (voir Reed, 1978, entre autres). Ce comportement est difficile à modéliser avec les réseaux compétitifs, qui, tel que déjà mentionné, fixent l'appartenance catégorielle. L'ajout d'unités dans un système distribué permettrait au système d'explorer une nouvelle partie de l'espace multidimensionnel, sans nécessairement perdre la flexibilité de réponse propre aux participants dans les expériences de catégorisation.

# CONCLUSION

Dans la présente thèse, un nouveau modèle perceptivo-cognitif autonome, soit FEBAM, a été proposé. Ce modèle permet de reproduire simultanément des processus de différentiation des entrées perceptuelles, de catégorisation, ainsi que d'extraction d'un vocabulaire de composantes perceptuelles iconiques. Cette approche autonome a été étendue pour tenir compte de rétroactions externes fournies par l'environnement, un ajout simple inspiré des expériences de laboratoires en psychologie cognitive.

Au Chapitre 1, plusieurs problématiques d'apprentissage et de catégorisation perceptuels ont été établies. Pour débuter, on a souligné l'importance de considérer l'apprentissage perceptuel, au niveau objet, comme un processus de différentiation progressive permettant de rendre les représentations d'objets plus précises et moins confuses entre elles. Aussi, il a été proposé que ce processus d'apprentissage perceptuel devrait être le fruit d'une réduction dimensionnelle, et qu'il devrait pouvoir être effectué de façon autonome. Ainsi, sans aucune rétroaction externe, le système perceptivo-cognitif pourrait différencier les objets, en enrichissant sa connaissance des régularités statistiques de l'environnement.

Ensuite, différentes théories de catégorisation ont été présentées. On retient principalement de cette recension la division nécessaire du système perceptivo-cognitif en deux modules, soit un sous-système servant à la mémorisation des exemplaires, et un autre servant à la mise en place d'un paysage catégoriel. Aussi, on a présenté des preuves empiriques de l'existence d'un système de création de composantes perceptuelles. Finalement, les propositions de plusieurs théoriciens face au développement d'un vocabulaire de composantes ont été explorées. Ces derniers ont proposé qu'un modèle perceptivo-cognitif devrait tenir compte de processus ascendants de création de caractéristiques, basés principalement sur des architectures de réseaux de neurones.

Au Chapitre 2, plusieurs classes générales de réseaux de neurones ont été couvertes. Il a été conclu, suite à l'étude des caractéristiques de ces réseaux, qu'aucune classe seule ne

pouvait rendre compte adéquatement des caractéristiques et tâches prédéfinies. Ainsi, il a été proposé qu'en unifiant des postulats définitoires provenant de différentes classes, il serait possible d'accomplir la totalité des tâches désirées.

Au Chapitre 3, un nouveau modèle d'inspiration perceptivo-cognitive a été proposé. Ce modèle utilise les postulats suivants, provenant de diverses classes de réseaux : 1) architecture de base bidirectionnelle (inspiré de la classe des BAM) utilisant des matrices de poids asymétriques; 2) présence d'une boucle de rétroaction (classe des RAM); 3) possibilité d'effectuer de la réduction dimensionnelle grâce à une couche intermédiaire d'unités (classe des auto-encodeurs); 4) principe d'apprentissage hebbien/anti-hebbien basé sur les différences temporelles (tout comme NDRAM et le BHM); 5) règle de transmission permettant la création d'attracteurs à des positions autres qu'aux vertex d'un hypercube (règle de NDRAM, réutilisée avec le BHM).

Le Chapitre 4 a exploré les comportements de base du modèle. De façon principale, il a été montré que FEBAM pouvait accomplir l'extraction de caractéristiques et la réduction dimensionnelle, le développement d'une mémoire d'exemplaires parfaits, ainsi que la catégorisation et l'identification autonomes. FEBAM a été maintes fois comparé à NDRAM, modèle duquel il émane. On a montré que FEBAM permettait de nettes économies cognitives, lorsque comparé à une RAM. Ces comparaisons ont aussi permis de s'assurer que l'ajout d'une couche de compression à NDRAM n'avait pas un impact significatif sur les capacités héritées de ce dernier modèle. On a pu montrer, entre autres, que malgré la nécessité d'initialiser les poids de connexion de façon aléatoire, ce facteur n'entrave pas la réussite du rappel parfait, à condition que le nombre d'unités de compression soit suffisant. Aussi, il a été montré que FEBAM produisait moins d'attracteurs nuisibles que NDRAM.

Finalement, au Chapitre 5, on a étendu le modèle FEBAM au cas où l'on veut associer aux stimulations perceptuelles des réponses prédéterminées. Il a été montré que le modèle résultant FEBAM-RA, pouvait effectuer un rappel parfait des réponses. Aussi, FEBAM-RA reproduit deux effets empiriques de façon qualitative. Premièrement, l'utilisation de deux modules séparés , le premier effectuant le traitement perceptuel, et le deuxième effectuant l'association d'une réponse à la compression développée, permet de reproduire le classique effet de pré-exposition de Gibson-Walk. Aussi, on a montré que si le traitement au niveau

«objet » et au niveau « catégorie » sont effectués par des sous-systèmes séparés, on peut reproduire la difficulté relative d'une tâche d'identification et d'une tâche de catégorisation effectuées simultanément.

# BIBLIOGRAPHIE


Abdi, H., Valentin, D., & Edelman, B. (1998). Eigenfeatures as intermediate level representations: the case for PCA models. *Brain and Behavioral Sciences, 21,* 17-18.

Anderson, J.A. (1972). A simple neural model generating an interactive memory. *Mathematical Biosciences. 14,* 197-220.

Anderson, J.A. (1995) An introduction to neural networks. Cambridge, MA : MIT Press.

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409-429.

Anderson, J.A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review, 84,* 413-451.

Ashby, F.G. (1992). Multidimensional models of categorization. *In* F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ: Erlbaum.

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998) A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442-481.

Ashby, F.G., & Gott, R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory & Cognition, 14,* 33-53.

Ashby, F.G., & Lee, W.W. (1993). Perceptual variability as a fundamental axiom of perceptual science. *In* S.C. Masin (Ed.), *Foundations of perceptual theory* (pp. 369-399). Amsterdam: Elsevier Science Publishers B.V.

Ashby, F.G., & Waldron, E.M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6,* 363-378.

Aydin, A., & Pearce, J.M. (1994). Prototype effects in categorization by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 264-277.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. *In* W.A. Rosenblinth (Ed.), *Sensory Communication* (pp.217-234). Cambridge: MIT Press.

Barlow, H.B. (1989). Unsupervised learning. *Neural Computation, 1,* 295-311.

Bégin, J., & Proulx, R. (1996). Categorization in unsupervised neural networks: the Eidos model. *IEEE Transactions on Neural Networks, 7,* 147-154.

Bénard, J., Stach, S., & Giurfa M. (2006). Categorization of visual stimuli in the honeybee apis mellifera. *Animal Cognition, 9,* 237-270,

Blair, M., & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition, 29,* 1153-1164.

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. *Memory & Cognition, 31,* 1293-1301.

Bogacz, R., Brown, M.W., & Giraud-Carrier, C. (2000). Frequency-based error back-propagation in a cortical network". *Proceedings of the 2000 International Joint Conference on Neural Networks,* 211–216.

Bomba, P.C., & Siqueland, E.R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology, 35,* 294–328.

Bourne, L.E. (1970). Knowing and using concepts. *Psychological Review, 77,* 546-556.

Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. *In* E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

Brooks, L.R. (1987). Non-analytic cognition. *In* U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual bases of categories.* Cambridge: Cambridge University Press.

Brooks, L.R., Norman, G.R., & Allen, S.W. (1991). The role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120,* 278-287.

Brown, T.H., Kairiss, E.W., & Keenan, C.L. (1990). Hebbian synapses–biophysical mechanisms and algorithms. *Annual Review of Neuroscience, 13,* 475-511.

Bruner, J., Goodnow, J., & Austin, A. (1956). *A Study of Thinking.* New York: Wiley.

Bruner, J.S., Wallach, M.A., & Galanter, E.H. (1959). The identification of recurrent regularity. *American Journal of Psychology, 72,* 200-209.

Carpenter, G.A., Grossberg, S., & Reynolds, J.H. (1991), ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, *Neural Networks, 4,* 565-588.

Chartier, S., & Boukadoum, M. (2006). A bidirectional heteroassociative memory for binary and grey-level patterns. *IEEE Transactions on Neural Networks, 17,* 385-396.

Chartier, S., Giguère, G., Renaud, P., Proulx, R. & Lina, J.-M. (2007). FEBAM: A feature-extracting bidirectional associative memory. *Proceedings of the 2007 International Joint Conference on Neural Networks,* 1679-1687.

Chartier, S., & Proulx, R. (2005). NDRAM: nonlinear dynamic recurrent associative memory for learning bipolar and nonbipolar correlated patterns. *IEEE Transactions on Neural Networks, 16,* 1393-1400.

Christos, G. A. (1996). Investigation of the Crik-Mitchison reverse-learning dream sleep hypothesis in dynamical settings. *Neural Networks, 9,* 427-434.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cousineau, D., Lacroix, G.L., & Hélie, S. (2003). Redefining the rules: Providing race models with a connectionist learning rule. *Connection Science, 15,* 27-43.

Diamantaras, K.I. & Kung, S.Y. (1996). *Principal component neural networks: theory and applications.* New York: John Wiley & Sons.

Diederich, S., & Opper, M. (1987). Learning of correlated pattern in spin-glass networks by local learning rules. *Physical Review Letters, 58,* 949–952.

Edelman, S., & Intrator, N. (1997). Learning as extraction of low-dimensional representations. *In* R.L. Goldstone, D.L. Medin, & P.G. Schyns (Eds.), *Perceptual learning: The psychology of learning and motivation, Vol. 36* (pp. 353–380). San Diego, CA: Academic Press.

Erickson, M.A. & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107-140.

Földiák, P. (1989). Adaptive network for optimal linear feature extraction. *Proceedings of the 1989 International Joint Conference on Neural Networks,* 401-405.

Freeman, J.A. (1994). *Simulating neural networks with Mathematica.* Reading, MA : Addison Wesley.

Garner, W.R. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Garnett, R., Huegerich, T., Chui, C.K., & He, W. (2005). A universal noise removal algorithm with impulse detector. *IEEE Transactions on Image Processing, 14,* 1747-1754.

Gerganov, A., Grinberg, M., Quinn, P.C., & Goldstone, R.L. (2007). Simulating conceptually-guided perceptual learning. *In* D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 287-292). Mahwah, NJ: Lawrence Erlbaum.

Gibson, E.J. (1963). Perceptual learning. *Annual Review of Psychology, 14,* 29-56.

Gibson, E.J. (1969). *Principles of perceptual learning and development.* New York : Meredith Corporation.

Gibson, J.J., & Gibson, E. (1955). Perceptual learning: differentiation or enrichment? *Psychological Review, 62,* 32-41.

Gibson, J.J., & Gibson, E. (1955). What is learned in perceptual learning? A reply to Prof. Postman. *Psychological Review, 62,* 447-450.

Gibson, E. J., & Walk, R.D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49,* 239-242.

Giguère, G., Chartier, S., Proulx, R., & Lina, J.M. (2007a). Category development and reorganization using a bidirectional associative memory-inspired architecture. *In* R.L. Lewis, T.A. Polk, & J.E. Laird (Eds.), *Proceedings of the 8th International Conference on Cognitive Modeling, (pp.97-102).* Ann Arbor, MI: University of Michigan.

Giguère, G., Chartier, S., Proulx, R., & Lina, J.M. (2007b). Creating perceptual features using a BAM-inspired architecture. *In* D.S. McNamara & J.G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society, (pp. 1025-1030).* Austin, TX: Cognitive Science Society.

Glanzer, M., & Cunitz, A.R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Behavior, 5,* 351–360.

Goldstein, E.B. (2008). *Cognitive psychology : connecting mind, research, and everyday experience.* Belmont, CA: Wadsworth.

Goldstone, R.L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585-612.

Goldstone, R.L., & Kersten, A. (2003). Concepts and categories. *In* A.F. Healy & R.W. Proctor (Eds.), *Comprehensive handbook of psychology, Volume 4: Experimental psychology* (pp. 591-621). New York: Wiley.

Goldstone, R.L., Schyns, P.G., & Medin, D.L. (1997). Learning to bridge between perception and cognition. *In* R.L. Goldstone, D.L. Medin & P.G. Schyns (Eds.), *Perceptual learning.* (pp. 1–14). San Diego, CA: Academic Press.

Grossberg, S. (1987), Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23-63.

Grossberg, S. (1988). *Neural networks and natural intelligence.* Cambridge, MA: MIT Press.

Hall, G. (1991). *Perceptual and associative learning.* Oxford : Clarendon Press.

Hamann, S.B., & Squire, L.R. (1997). Intact priming for new perceptual representations in amnesia. *Journal of Cognitive Neuroscience, 9,* 699-713.

Harnad, S. (1990). The symbol grounding problem. *Physica D., 42,* 335-346.

Harnad, S. (2005). To cognize is to categorize: cognition is categorization. *In* C. Lefebvre & H. Cohen, (Eds). *Handbook of categorization in cognitive science* (pp.20-45). Oxford : Elsevier.

Hassoun, M.H. (1989) Dynamic heteroassociative neural memories. *Neural Networks, 2,* 275-287.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16,* 321-338.

Haykin, S. (1994). *Neural networks: a comprehensive foundation.* New York : Cambridge University Press.

Hélie, S. (2008). Energy minimization in the nonlinear dynamic recurrent associative memory. *Neural Networks, 21,* 1041-1044.

Hélie, S., Chartier, S., & Proulx, R. (2006). Are unsupervised neural networks ignorant? Sizing the effect of environmental distributions on unsupervised learning. *Cognitive Systems Research, 7,* 357-371

Hertz, J., Krogh, A., & Palmer, R.G. (1991). *Introduction to the theory of neural computation.* Redwood City, CA : Addison-Wesley.

Hinton, G.E., & Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313,* 504-507.

Homa, D. (1978) Abstraction of ill-defined form. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 407–416.

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 11-23.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418-439.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA, 9,* 2554-2558.

Hopfield, J.J., Feinstein, D.I., & Palmer, R.G. (1983). Unlearning has a stabilizing effect in collective memories. *Nature, 304,* 158-159.

Hornik K., Stinchcombe M., & White H. (1989): Multilayer feed-forward networks are universal approximators, *Neural Networks, 2,* 359-366.

Hyvärinen, A., Hurri, J., & Väyrynen, J. (2003). Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A, 20,* 1237-1252.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks, 13,* 411-430.

Ittelson, W.H. (1962). Perception and transactional psychology. *In* S. Koch (Ed.), *Psychology : a study of science* (pp. 660-704). New York : McGraw-Hill.

Kanter, J., & Sompolinsky, H. (1987). Associative recall of memory without errors. *Physical Review A, 35,* 380–392.

Karhunen, J., Pajunen, P., & Oja, E. (1998). The nonlinear PCA criterion in blind source separation: relations with other approaches. *Neurocomputing, 22,* 5-20.

Knapp, A. G., & Anderson, J. A. (1984), Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning. Memory, & Cognition, 10,* 616-631.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273,* 1399-1402.

Knowlton, B.J., & Squire, L.R. (1993). The learning of categories: Parallel brain systemsfor item memory and category knowledge. *Science, 262,* 1747–1749.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers, C-21,* 353-359.

Kohonen, T. (1989). *Self-organization and associative memory.* Berlin : Springer-Verlag.

Komatsu, L. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112,* 500-526.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics, 18,* 49-60.

Kosko, B. (1990). Unsupervised learning in noise. *IEEE Transactions on Neural Networks, 1,* 44-57.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kruschke, J.K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22,* 3-26.

Lacroix, G.L., Giguère, G., & Larochelle, S. (2005). The origin of exemplar effects in rule-driven categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31,* 272-288.

Leung, C.S. (1994). Optimum learning for bidirectional associative memory in the sense of capacity. *IEEE Transactions on Systems, Man, and Cybernetics, 24,* 791-796.

Logan, G.D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492-527.

Marsolek, C.J. (1995). Abstract visual-form representations in the left cerebral hemisphere. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 375-386.

Marsolek, C.J., Kosslyn, S.M., & Squire, L.R. (1992). Form-specific visual priming in the right cerebral hemisphere. *Journal of Experimental Psychology: Learning, Memory & Cognition, 18* , 492-508.

McCloskey, M.E., & Glucksberg, S. (1978) Natural Categories: well defined or fuzzy sets? *Memory & Cognition, 6,* 462-472

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469-1481.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory, 7,* 355-368.

Mervis, C., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society, 7,* 283-284.

Minda, J.P., & Smith, J.D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 28,* 275-292.

Murphy, G.L. (2002). *The big book of concepts.* Cambridge : MIT Press.

Murphy, G.L., & Brownell, H.H. (1985). Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11,* 70-84.

Negnevitsky, M. (2001). *Artificial intelligence: a guide to intelligent systems.* Boston : Addison-Wesley.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 104-114.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Nosofsky, R.M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory & Cognition, 14,* 54-65.

Nosofsky, R.M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. *In* F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363-393). Hillsdale, NJ: Lawrence Erlbaum.

Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266-300.

Nosofsky, R.M., & Zaki, S.R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: an exemplar-based interpretation. *Psychological Science, 9,* 247-255.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15,* 267–273

Oja, E. (1989). Neural networks, principal components and subspaces. *International Journal of Neural Systems, 1,* 61-68.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinions in Neurobiology, 14,* 481-487.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2,* 455-462.

Palmeri, T.J., & Flanery, M.A. (1999). Learning about categories in the absence of training: profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science, 10,* 526-530

Perrin, N. (1992). Uniting identification, similarity, and preference : general recogntion theory. *In* F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 123-146). Hillsdale, NJ: Erlbaum.

Personnaz, L., Guyon, I., & Dreyfus, G. (1985). Information storage and retrieval in spin-like neural networks. *Journal de Physique-Lettres, 46,* L359-L365

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353-363.

Posner, M.I., & Keele, S.W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304-308.

Rainer, G., & Miller, E.K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron, 27,* 179–189.

Raymond, J.E., Shapiro, K.L, & Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18,* 849-860.

Reber, P.J., Stark, C.E.L, & Squire, L.R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences of the USA*, 95, 747-750

Reber, P.J., Stark, C.E.L, & Squire, L.R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning and Memory*, 5, 420-428

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 383-407.

Reed, S.K. (1978). Category vs. item learning: Implications for categorization models. *Memory & Cognition, 6,* 612–621.

Reed, M.J., Hamann, S.B., Stefanacci, L., & Squire, L.R. (1997). When amnesic patients perform well on recognition memory tests. *Behavioral Neuroscience, 111,* 1163-1170.

Rips, L.J., Shoben, E.J., Smith, E.E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1-20.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4,* 328-350.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E. (1978). Principles of categorization. *In* E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp.27-48).Hillsdale, NJ : Erlbaum.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605

Rosch, E., Mervis, C.B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Rosch, E., Simpson, C., & Miller, R.S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 491-502.

Rubner, J., & Schulten, K. (1990). Development of feature detectors by self-organization: a network model. *Biological Cybernetics, 62,* 193-199.

Rumelhart, D.E. and Zipser, D. (1986). Feature discovery by competitive learning. *In* D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of cognition, Volume 1: Foundations* (pp. 151-193). Cambridge, MA : MIT Press.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. *In* D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of cognition, Volume 1: Foundations* (pp. 318-364). Cambridge, MA : MIT Press.

Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks, 2,* 459-473.

Schacter, D.L. (1994). Priming and multiple memory systems: perceptual mechanisms of implicit memory. *In* D.L. Schacter & E. Tulving (Eds.), *Memory systems* (pp. 244–256). Cambridge, MA : MIT Press.

Schyns, P.G., Goldstone, R.L. & Thibaut, J.P. (1998). The development of features in object concepts. *Behavioral & Brain Sciences, 21,* 1-54.

Schyns, P.G., & Murphy, G.L. (1991). The ontogeny of units in object categories. *Proceedings of the 13th Meeting of the Cognitive Science Society,* 197-202.

Schyns, P.G., & Murphy, G.L. (1994). The ontogeny of part representation in object concepts. *In* D.L. Medin (Ed.), *The psychology of learning and motivation : Volume 31* (pp. 305-354). San Diego, CA: Academic Press.

Schyns, P.G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23,* 681-696.

Shepard, R.N. (1958). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325-345.

Shepard, R.N. (1958). Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55,* 509-523.

Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika, 27,* 125-140, 219-246.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Shin, H.J., & Nosofsky, R.M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121,* 278-304.

Smith, E.E., & Medin, D.M. (1981). *Categories and concepts.* Cambridge: Harvard University Press.

Smith, J.D. (2005). Wanted : a new psychology of exemplars. *Canadian Journal of Experimental Psychology, 59,* 47-53.

Smith, J.D., & Minda, J.P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1411-1436.

Smith, J.D., & Minda, J.P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory &Cognition, 26,* 3-27.

Smith, J.D., & Minda, J.P. (2001). Journey to the center of the category: The dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27,* 9841002.

Smith, J.D., & Minda, J.P. (2002). Distinguishing prototype-based and exemplar-based processes in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 800-811

Smith, J.D., Murray, M.J., & Minda, J.P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory & Cognition, 23,* 659-680.

Smith, J.D., Redford, J.S., & Haas ,S.M. (2008). Prototype abstraction by monkeys (Macaca mulatta). *Journal of Experimental Psychology : General, 137,* 390-401.

Squire, L.R., & Knowlton, B.J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences, USA, 92,* 12470–12474.

Sperling, G. (1960). The information available in brief visual presentations. *Psychology Monographs, 74,* 498.

Stork, G. (1989). Is backpropagation biologically plausible? *Proceedings of the 1989 International Joint Conference on Neural Networks,* 241-246.

Storkey, A.J., & Valabrègue, R. (1999). The basins of attraction of a new Hopfield learning rule. *Neural Networks, 12,* 869-876.

Sutton, R.S. (1988). Learning to predict by the methods of temporal difference. *Machine Learning, 3,* 9-44.

Thorpe, S. J., O'Regan, J. K., & Pouget, A. (1989). Humans fail on XOR pattern classification problems. *In* L. Personnaz & G. Dreyfus (Eds.), *Neural networks : from models to applications* (pp. 12-25). Paris: I.D.S.E.T.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17,* 401-419.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207-232.

Wang, Z. (1996). A bidirectional associative memory based on optimal linear associative memory. *IEEE Transactions on Computers, 10,* 1171-1179.

Wattenmaker, W.D., Dewey, G.I., Murphy, T.D., & Medin, D.L. (1986). Linear separability and concept learning: context, relational properties, and concept naturalness. *Cognitive Psychology, 18,* 158–194.

Wills, A.J., & McLaren, I.P.L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology, 51B,* 235–270.

Wittgenstein, L. (1953). Philosophical investigations, sections 65-78. *In* E. Margolis & S. Laurence (Eds.), *Concepts : Core readings* (pp.171-175). Cambridge : MIT Press.

Zhang, L., & Mei, J. (2003). Shaping up simple cell's receptive field of animal vision by ICA and its application in navigation system, *Neural Networks, 16,* 609-615.