

Aperçu des hypothèses computationnelles utilisées pour modéliser la production du langage

F.-Xavier Alario

*CNRS & Université de Provence (UMR 6146)
Marseille*

F.-Xavier ALARIO
Laboratoire de Psychologie Cognitive
CNRS & Université de Provence
3, place Victor Hugo - Case 66
13331 Marseille Cedex 3, France

04 91 10 67 91

alario@up.univ-mrs.fr

Résumé

Nous passons en revue les hypothèses computationnelles qui ont pu être utilisées pour rendre compte des processus psycholinguistiques impliqués dans la production du langage. Le chapitre commence par une section décrivant l'architecture générale des réseaux de neurones artificiels couramment employés dans ce domaine de recherche. Cette section décrit les hypothèses de base partagées par de nombreux modèles comme sont la structure du réseau, la nature du codage de l'information et les équations de propagation d'activation. La deuxième section est centrée sur le processus de sélection lexicale — choix du mot approprié pour exprimer un message donné. Cette section commence par une brève présentation du processus psychologique à modéliser. Elle comprend ensuite des analyses de la nature de la propagation d'activation (de séquentielle à interactive) dans différents modèles de production, de l'implémentation du critère de sélection des unités formant la sortie du modèle et, pour finir, de la modélisation des lésions neuropsychologiques qui affectent directement le processus de sélection lexicale.

Il est conclu qu'à ce stade de la recherche les modèles doivent plutôt être vus comme des outils d'exploration d'hypothèses que comme des reflets fidèles des processus psychologiques.

Dans ce chapitre nous décrivons quelques unes des hypothèses computationnelles qui ont été utilisées pour formuler des modèles implémentés de la production du langage. Autrement dit, nous nous intéressons à la façon dont les hypothèses cognitives sont formulées en termes mathématiques lors de la construction des modèles. Notre objectif est de proposer une introduction descriptive et méthodologique. Nous ne chercherons pas à formuler de nouvelle proposition théorique, ni à évaluer la validité de certains modèles ou hypothèses en regard de données empiriques.

Les modèles computationnels de la production du langage sont des constructions complexes et variées. Ils représentent les informations et les processus à l'aide d'hypothèses diverses. De plus, ils ne se donnent pas tous le même objectif : tous ne cherchent pas à rendre compte du même genre de données (p.ex., simulation de patterns d'erreurs ou bien de temps de réponse), ni de la même façon (p.ex., simulation simple d'effets globaux ou des distributions complètes produites par les sujets). Dans ce contexte, une analyse spécifique de l'implémentation computationnelle des hypothèses paraît utile pour plusieurs raisons. D'une part, elle peut contribuer à décortiquer les modèles proposés et leurs mécanismes. On peut ainsi espérer que les raisons pour lesquelles – par exemple – tel effet est simulé ou pas pourront être plus facilement mises à jour. D'autre part, une meilleure appréhension de la signification de la formulation mathématique permet de vérifier son adéquation à l'hypothèse cognitive qu'elle est censée représenter. Finalement, cette analyse constitue une étape essentielle pour la comparaison de modèles formulés à l'aide de formalismes qui ne sont pas identiques. A ce propos, on pourra constater ci-dessous la diversité des détails d'implémentation qui ont pu être utilisés par différents auteurs.

Au sein de la psycholinguistique, l'étude de la production du langage s'intéresse aux processus permettant de récupérer en mémoire l'information linguistique pour exprimer un message donné. Le principal dénominateur commun de ce champ d'étude est le fait que les traitements linguistiques étudiés sont, à la base, sous contrôle d'informations sémantiques ou conceptuelles (Levelt, 1989). Une telle définition exclut de fait certaines situations où il peut pourtant y avoir production de langage, comme par exemple la lecture de mots. En effet, dans ce type de comportement les processus étudiés sont sous contrôle d'informations linguistiques – les représentations orthographiques – plutôt que d'informations sémantiques ou conceptuelles. Même si des domaines comme la lecture de mots ont

bénéficié d'une importante activité de modélisation (ex. Ans, Carbonnel, & Valdois, 1998; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Plaut, McClelland, Seidenberg, & Patterson, 1996; Zorzi, Houghton, & Butterworth, 1998), nous limiterons notre discussion aux choix d'implémentation utilisés dans les modèles de production de langage¹, et cela pour deux raisons. Au fil des années s'est constitué, au sein de la psycholinguistique, une unité de la recherche dans le domaine ainsi défini (Wheeldon, 2000). De plus, les modèles computationnels de production de langage sont relativement peu nombreux, ce qui rend abordable le type d'analyse que nous proposons ici.

Nous considérerons des modèles proposés pour rendre compte des processus de production de langage en nous intéressant aux processus de production de mots isolés. Dans la suite du chapitre, nous commencerons par présenter les principes très généraux de la modélisation computationnelle. Ensuite nous passerons en revue les principales implémentations utilisées pour rendre compte de la sélection lexicale –choix du mot approprié au message à exprimer– qui est le processus central de nombreux modèles de production. Essentiellement pour des raisons de place, nous nous verrons obligés de laisser de côté d'autres aspects importants des traitements de production, comme l'encodage phonologique (Dell, Burger, & Svec, 1997; Dell & O'Seaghdha, 1994) ou la construction de structures syntaxiques.

1) Les éléments constitutifs des modèles

1.1) Structure d'un réseau de neurones

Dans la très grande majorité des cas, les modèles computationnels des processus de production de mots sont basés sur des ensembles d'unités de représentation (neurones artificiels). D'un point de vue mathématique, ces unités sont des suites numériques dont l'indice est une dimension représentant le temps (discret) dans lequel évolue le modèle. Ces unités représentent des informations pertinentes pour la réalisation du comportement étudié – par exemple des informations linguistiques, mais aussi des informations non directement interprétables psychologiquement. Les valeurs prises au cours du temps par la suite numérique correspondant à une unité donnée constituent le degré d'activation de

¹ Bien entendu, nombre de principes de modélisation des processus de lecture et d'autres comportements se retrouvent dans la modélisation des processus de production (et inversement).

l'information que l'unité représente. Ces niveaux d'activation sont liés entre eux par un système d'équations qui régissent leur évolution en fonction de la dimension temporelle. Dans le cas général, la valeur de l'activation d'une unité à un temps donné est une combinaison des valeurs des autres suites au temps précédent et possiblement d'autres paramètres. Les équations qui lient les activations des unités entre elles sont une traduction des hypothèses cognitives que l'on cherche à implémenter. Sur la base de ce système d'équations et d'un ensemble de valeurs initiales, on peut suivre l'évolution des « activations » de chacune des informations en fonction du « temps » (*cf.* Figure 1).

< Figure 1 ici >

Les valeurs d'activation permettent de définir les réponses « produites » par le modèle. Cette définition est le plus souvent basée sur des hypothèses supplémentaires impliquant la notion de sélection (p.ex., quelles unités sont pertinentes ? à quelles conditions faut-il les considérer ? etc.) Selon le critère de sélection implémenté on pourra modéliser l'une ou l'autre des données généralement observées dans l'étude de la production du langage : les erreurs de production et les latences des réponses.

Certains modèles sont définis comme des fonctions déterministes alors que d'autres sont définis comme des variables aléatoires. Dans les premiers un état initial donné conduit toujours, par la même trajectoire, au même état final. Dans le deuxième cas, l'introduction de variables aléatoires conduit, pour chaque état initial, à une distribution probabilisée d'états finaux possibles. En termes psycholinguistiques, cette deuxième famille de modèles a l'avantage de pouvoir simuler, par exemple, des variations interindividuelles, inter items, ou de simuler l'ensemble de la distribution de réponses plutôt qu'une valeur représentant sa moyenne.

1.2) La nature des représentations

La représentation d'information dans les unités du modèle peut se faire de façon localiste (ex. Dell, 1986) ou distribuée (ex. Lambon Ralph, McClelland, Patterson, Galton, & Hodges, 2001). L'interprétation localiste attribue à chaque unité –ou, tout au moins, à la plupart d'entre elles– la valeur de l'activation d'une information linguistiquement ou psychologiquement définie. Par exemple, les valeurs prises au cours du temps par l'une des unités du réseau (que nous appellerons l'unité B13)

représenteront l'activation en mémoire de l'entrée lexicale du mot *différence* (Page, 2000). Dans une interprétation distribuée, l'activation d'une information linguistique –par exemple, l'entrée lexicale de *différence*– est prise en charge par un ensemble d'unités du réseau. Ces unités coderont en fait tout un ensemble d'informations linguistiques. Par exemple, les unités appelés B01, B02, B03, ... à B20 pourront représenter, ensemble, les mots « connus » par le système. Chaque mot connu (p.ex., *différence*) est alors représenté par une collection prédéfinie de vingt valeurs d'activation que les unités B01 à B20 pourront prendre. Pour un t donné, plus la valeur du vecteur [B01(t) ... B20(t)] sera proche (au sens d'une distance à définir) des valeurs spécifiées pour *différence*, plus cette entrée lexicale sera activée en mémoire (Hinton, McClelland, & Rumelhart, 1986).

La plupart des modèles computationnels de production de mots postulent une représentation localiste de l'information, avec quelques exceptions (modèles de sélection lexicale de Lambon Ralph et al., 2001, ou de Plaut & Shallice, 1993 ; modèle d'encodage phonologique de Dell, Juliano, & Govindjee, 1993). On peut penser que les modèles distribués ne sont pas souvent utilisés car ils sont intrinsèquement moins lisibles - le codage des informations y est moins transparent que dans un modèle localiste. De plus, les avantages souvent mis en avant pour motiver l'utilisation d'un modèle distribué - apprentissage, généralisation, inférence sur la base d'information partielle - ne sont pas généralement pas ceux que l'on cherche à modéliser en production du langage. Dell et collaborateurs (1993) ont utilisé une représentation distribuée des informations phonologiques pour argumenter que l'encodage à ce niveau pouvait se passer des représentations structurelles généralement postulées.

Les modèles localistes et distribués implémentent les hypothèses psychologiques de façon assez différente. On peut comparer ces deux familles en considérant les représentations localistes comme des représentations distribuées fortement pondérées (Figure 2, gauche). Dans un modèle localiste on considérera généralement que l'entrée lexicale de *différence* est significativement activée si (par exemple) l'unité B13 a une valeur d'activation beaucoup plus forte que les unités représentant les autres entrées lexicales (le détail de cette affirmation dépend bien sûr du critère de sélection postulé). Une interprétation de ce codage en termes distribués consiste à dire que la collection de valeurs représentant *différence* dans le modèle est proche d'un vecteur avec des valeurs très faibles à toutes ses coordonnées sauf à la 13^{ème}. Cette contrainte *a priori* sur la représentation des informations n'est bien

sûr pas présente dans un modèle distribué où les représentations et celles-ci peuvent impliquer des valeurs élevées pour l'activation de plusieurs unités, très faibles pour d'autres, etc. (cf. Figure 2, droite). D'autres types de contraintes s'appliquent généralement aux modèles localistes. Ceux-ci postulent une unité dans le réseau pour chaque unité d'information à représenter là où les modèles distribués postulent beaucoup plus de suites (d'unités) qu'il n'y a d'unités informationnelles (psychologiques) à prendre en compte. Le postulat de représentations localistes va souvent de pair avec des limitations sur la connectivité dans le modèle (cf. ci-dessous) qui ne sont pas toujours imposées dans les modèles distribués.

< Figure 2 ici >

En somme, les représentations localistes sont équivalentes aux représentations distribuées si on y adjoint quelques contraintes sur l'interprétation des séries numériques et le codage de l'information. Inversement, il est possible dans certains cas qu'un modèle distribué puisse être reformulé en termes localistes, si la structure du codage de l'information s'y prête. Un tel point de vue est utile pour comparer les mérites de différents modèles.

1.3) La connectivité dans le modèle

Comme cela a été dit plus haut, un modèle comprend un ensemble d'unités dont les valeurs d'activation dépendent généralement des valeurs d'activation d'autres unités du modèle. Deux unités dont les valeurs sont dépendantes l'une de l'autre sont dites connectées entre elles : l'activation peut se propager entre les informations qu'elles représentent. Si la relation entre l'activation de deux unités est croissante, on dit que la relation est d'activation ; si elle est décroissante, on parle d'inhibition. La formule générale de propagation d'activation utilisée couramment *dans les modèles de production de langage* (ainsi que dans de nombreux réseaux de neuronaux) est

$$\text{activation}_i(t) = [\text{activation}_i(t-1) \times d] + \sum_j [\text{activation}_j(t-1) \times p_{ij}] + \text{bruit}$$

où i : indice de l'unité courante

j : indice parcourant l'ensemble des unités connectées à l'unité i

d : paramètre d'atténuation de l'activation

p_{ij} : paramètre de connexion entre les unités i et j .

L'activation d'une unité est ainsi dépendante de sa propre activation et de celle des unités auxquelles elle est connectée. Un paramètre important est le poids relatif de ces deux contributions dans l'activation des unités. On peut le rendre explicite en considérant le rapport entre la contribution à l'activation due à l'unité elle-même (d) et la contribution à l'activation due aux autres unités ($p \times m$, où p est la valeur moyenne du paramètre de connexion et où m est le nombre moyen de connexions arrivant à une unité). On constate qu'une unité résistera plus au changement (elle aura plus « d'inertie ») si $d > (p \times m)$ et qu'inversement une unité sera plus sensible à son contexte dans le modèle dans le cas contraire. Par ailleurs, on peut approximativement affirmer que si $d + (p \times m) > 1$ alors les activations des unités vont croître de façon indéfinie alors que si $d + (p \times m) < 1$ elles finiront par décroître vers zéro. Explorer ces rapports dans les différentes parties d'un modèle spécifié numériquement aide à en comprendre de façon détaillée l'évolution temporelle. De fait, les études de lésions de modèles s'intéressent à la façon dont des modifications de ces paramètres peuvent affecter son fonctionnement (*cf.* section sur ce sujet ci-dessous).

Les variantes de la formule donnée ci-dessus portent généralement sur (a) l'utilisation possible de fonctions non linéaires modifiant l'accumulation d'activation et ses limites possibles (ex. fonction sigmoïde modulant l'activation pour la borner entre des valeurs finies ; Starreveld & La Heij, 1996), et (b) sur quelles unités peuvent contribuer à l'activation d'une unité donnée – sa connectivité. Le détail de la spécification de la connectivité permet de définir des familles d'unités ayant des patterns de connectivité similaires, et qui représenteront des informations (linguistiques) similaires. On définit par là même différents niveaux de traitement. Dans de nombreux modèles de production de langage, les unités au sein d'un niveau de traitement ne sont généralement pas connectées entre elles : elles n'ont de connexions que vers les unités de niveaux inférieurs, ou supérieurs (sauf parfois dans les niveaux sémantiques, ou dans quelques modèles comme Harley, 1993; Laine, Tikkala, & Juhola, 1998; ou Lambon Ralph et al., 2001). Cette absence de connexions latérales (i.e., au sein d'un même niveau) contraste avec la pratique courante dans d'autres champs d'investigation, par exemple la lecture de mots, où ce genre de connexion est souvent postulé (*cf.* modèles cités précédemment). Les connexions latérales, généralement inhibitrices, ont pour rôle computationnel d'augmenter la différence d'activation entre unités. Dans un modèle à connexions latérales inhibitrices, une unité ayant une

activation plus élevée que les autres membres de son niveau verra cette différence croître au cours du temps (citation connexions latérales). Sans ces connexions, les différences d'activation ne sont pas démultipliées. On peut donc considérer que le choix d'utiliser des connexions latérales est étroitement lié au choix du processus de sélection qui sera implémenté, selon qu'il nécessite des différences d'activation fortes entre les unités ou pas. Comme on le verra ci-dessus, les modèles de production de langage mettent un accent important sur la définition du mécanisme de sélection.

2) Processus de sélection lexicale

Dans cette section nous passons en revue les hypothèses portant spécifiquement sur le processus de sélection lexicale qui assure le choix du mot approprié au message à exprimer parmi tous ceux qui sont connus du locuteur.

2.1) La nature de la propagation d'activation : de sérielle à interactive et retour

Comme nous avons vu, dans un modèle computationnel l'activation se propage d'une unité de représentation r1 à une unité de représentation r2 lorsque la valeur de l'activation de r2 est dépendante de la valeur de l'activation de r1 –typiquement, la valeur de l'activation de r1 à un temps donné est prise en compte dans le calcul de la valeur de r2 au temps suivant. La question de la connectivité occupe une place importante dans les débats en psychologie du langage. Pour décrire la formulation de cette question dans le domaine de la production, nous commençons par préciser quelques aspects du traitement qui ne sont pas controversés. Dans une situation de production (par exemple, dans la situation expérimentale où un locuteur doit dénommer un dessin) le système de production commence par activer des informations sémantiques correspondant au message à exprimer. Ces informations vont activer plusieurs candidats lexicaux parmi lesquels s'effectuera le choix du plus approprié. Ce candidat étant choisi, le système activera et récupérera les propriétés phonologiques correspondantes. Dans ce contexte, la question de la propagation d'activation dans le système de production a été formulée de la façon suivante. (1) Est-ce que les différentes entrées lexicales qui sont activés au cours du processus de sélection, mais qui pour autant ne seront pas produites (puisque'un seul candidat est choisi), activent leurs propriétés phonologiques ? Dans le cas d'une réponse affirmative on parlera de propagation d'activation en cascade et dans le cas contraire de propagation d'activation séquentielle. (2) Sous

l'hypothèse d'une propagation d'activation en cascade, est-ce que les informations phonologiques ainsi « pré-activées » peuvent influencer en retour le processus de sélection lexical ? Si oui, on parlera de rétroaction ou interactivité, dans le cas contraire de sérialité.

Quasiment toutes les réponses logiquement envisageables ont pu être proposées et défendues pour rendre compte de tel ou tel phénomène empirique : séquentialité stricte (Levelt, Roelofs, & Meyer, 1999), traitement en cascade sans rétroaction (Humphreys & Riddoch, 1988), interactivité bidirectionnelle (Dell, Schwartz, Martin, & Gagnon, 1997), etc. Chacune de ces études avançait un modèle de caractéristiques différentes, compatible avec un ensemble de données expérimentales. Toutefois, ils n'évaluaient pas toujours systématiquement la capacité des variantes de ces modèles à rendre compte des données. Ce point est important. Par exemple, Harley (1993) a pu montrer que des données impliquant en apparence un traitement séquentiel pouvaient aussi être expliquées dans un modèle en cascade. La logique de cette étude était différente des précédentes. Plutôt que de chercher à produire un modèle qui rende compte d'une série de données, cet auteur cherchait à étudier les propriétés générales d'un ensemble d'hypothèses computationnelles pour rendre compte des données.

Suivant cette même logique, Rapp et Goldrick (2000) ont proposé une évaluation très systématique de différentes hypothèses possibles concernant la propagation d'activation (voir aussi Dell, Martin, Saffran, Schwartz, & Gagnon, 2000; Dell, Schwartz et al., 1997; Foygel & Dell, 2000; Harley, 1993; Rumel & Caramazza, 2000; Rumel, Caramazza, Shelton, & Chialant, 2000). Pour cela, ils forment la dichotomie de sérialité–interactivité comme un continuum à plusieurs dimensions. Au sein d'un modèle simple de propagation d'activation à quatre niveaux, ils manipulent de façon multiple la quantité d'information que les représentations peuvent échanger. Leurs manipulations concernent : le coefficient qui module l'échange d'activation dans les équations du modèle, le sens (en avant donc sériel ou bien en avant et arrière donc interactif) de cet échange, le niveau du modèle où cette dimension est modulée et la quantité de bruit présente dans le système. Ils manipulent aussi l'impact d'une sélection sur le niveau d'activation de la représentation sélectionnée. Cette dimension affecte le caractère plus ou moins séquentiel du traitement. En effet, plus une sélection a d'impact en termes de niveau d'activation (ex. si la sélection est modélisée par une sur-activation de valeur 25 plutôt que de valeur 5), plus cette sélection créera une discontinuité dans les fonctions d'activation des

unités en aval du niveau où la sélection a lieu. Cette discontinuité peut être interprétée comme marquant le passage d'un certain mode de traitement (ex. traitement lexical) à un autre mode de traitement au niveau suivant (ex. phonologique), puisque ce niveau suivant reçoit, au moment de la sélection, une activation supplémentaire importante. Ainsi, une sélection plus forte équivaut à un traitement plus séquentiel (Dell & O'Seaghdha, 1991).

Au total, ces différentes manipulations permettent d'obtenir une activation plus ou moins en cascade, ainsi qu'une rétroaction et une sérialité plus ou moins fortes aux différents niveaux du modèle. Cette approche illustre la possibilité d'utiliser plusieurs hypothèses computationnelles qui ne sont pas équivalentes entre elles pour implémenter une hypothèse psychologique donnée (à savoir : le degré d'interactivité). Le paramétrage proposé permet de définir un ensemble de modèles variant selon le degré d'interactivité qu'ils intègrent et d'explorer leurs validités respectives. Le choix du meilleur modèle correspond alors à l'optimisation des paramètres du continuum afin de rendre compte au mieux des données à simuler. Une étude comme celle de Rapp et Goldrick (2000) contribue à établir quel degré d'interactivité est véritablement nécessaire dans un système en même temps qu'elle indique pourquoi d'autres degrés d'interactivité (p.ex. sérialité pure) ne conviendrait pas pour rendre compte du set de données qui est exploré. L'exploration exhaustive de l'espace des modèles possibles permet de mieux comprendre les raisons pour les caractéristiques des variantes du modèle.

Clairement, l'avantage de l'approche de Rapp et Goldrick (2000) est que l'espace des modèles possibles est spécifié de façon explicite et qu'il est exploré exhaustivement. Une limitation qui lui est inhérente est toutefois que la recherche se limite à cet espace. Dans cette approche, on cherche nécessairement à rendre compte des observations empiriques en localisant un modèle dans le continuum d'interactivité. Cela exclu *a priori* la possibilité d'imaginer d'autres mécanismes non stipulés dans les hypothèses de base (ex. mécanisme de monitoring des productions verbales, boucle externe de rétroaction) qui nécessiteraient une implémentation spécifique.

2.2) Critères de sélection d'unités

Sur la base de la connectivité définie dans le modèle, l'activation des unités va évoluer dans le temps. A un moment donné certaines unités seront sélectionnées pour constituer la réponse du modèle.

Il existe une grande diversité dans les implémentations du critère de sélection adoptées par les modèles de production de langage. Celles-ci sont dictées par le type de données que l'on cherche à modéliser, et par la structure des séries numériques. Dans la modélisation de la production d'erreurs, un état initial donné peut selon les conditions conduire à plusieurs états finaux, l'un d'entre eux étant considéré comme correct et les autres comme des erreurs de différents types. Dans la modélisation des latences de réponse, un état initial peut être associé au nombre de cycles nécessaires au modèle pour atteindre l'état final prédéfini correspondant.

Un premier critère de sélection consiste à postuler que la « réponse » se produit à un moment fixe. Par exemple, après un nombre fixe n de cycles l'unité sélectionnée est celle qui est la plus active (modèle localiste), ou bien le pattern global sélectionné est celui qui est obtenu après n cycles (modèle distribué ; Figure 3A). Cette méthode d'implémenter la sélection est possiblement la plus simple que l'on puisse imaginer, et l'une des plus fréquemment utilisées (ex. Dell, 1986; Dell, Schwartz et al., 1997; Foygel & Dell, 2000; Rapp & Goldrick, 2000; Rumel et al., 2000; Vousden, Brown, & Harley, 2000). Puisque le délai de sélection est fixe, ce type de critère de sélection permet de modéliser la précision des réponses mais pas leur délai². De plus ces modèles ne tiennent généralement pas compte du niveau d'activation auquel se trouvaient les unités lorsque la sélection a lieu (*cf.* plus bas pour des exceptions). Un modèle implémentant la sélection par critère temporel fixe conduira à des résultats satisfaisants si toutes les trajectoires issues des états initiaux s'ordonnent de façon appropriée –i.e., réponse attendue la plus active– en un nombre prédéterminé de cycles temporels. Autrement dit, lorsque les trajectoires issues des différents états initiaux croisent la ligne de seuil dans le bon ordre.

Une modélisation proche du critère de sélection temporel fixe consiste à modéliser la sélection sur la base d'un seuil absolu d'activation : une unité est sélectionnée si son activation dépasse une certaine valeur fixée à l'avance, quelle que soit la valeur de l'activation des autres unités à ce moment là (MacKay, 1987; Santiago, MacKay, Palma, & Rho, 2000). Ce type d'implémentation permet en principe de modéliser à la fois deux caractéristiques des réponses : leur latence –cycle du modèle où le seuil est dépassé– et leur précision –quelle série dépasse le seuil. En effet, contrairement à la sélection

² La valeur de n peut-être manipulée comme un paramètre, généralement interprété comme le débit de parole (n petit, locuteur plus rapide).

par seuil temporel on tient ici compte du niveau d'activation et on peut donner l'interprétation de « latence » au moment où le seuil a été atteint. Comme il a été dit pour les modèles à critère temporel fixe, un modèle à critère d'activation fixe produit des résultats satisfaisants si les fonctions d'activation atteignent de façon ordonnée le critère de sélection (figure 3B).

< Figure 3 ici >

En fait, dans certaines conditions ces deux critères de sélection sont formellement comparables, sinon équivalents. C'est particulièrement le cas si le comportement de l'activation des unités d'output (qu'elles soient considérées individuellement ou en groupe) est globalement croissant. Si tel est le cas, et étant donné un ensemble de trajectoires d'activation en fonction du temps, décider de la sélection en base à un critère temporel ou à un critère d'activation est équivalent, *même si les trajectoires des fonctions d'activation se croisent*. En d'autres termes : des trajectoires qui sont ordonnées à un temps pré spécifié pourront toujours être ordonnées de la même façon à un niveau d'activation pré spécifiée et le modèle produira les mêmes résultats avec les eux critères. Les fonctions d'activation croissantes sont en fait bijectives et que l'on peut échanger les rôles des dimensions d'activation et de temps³. Par contre, si les fonctions d'activation ne sont pas globalement croissantes (Figure 3C) alors il n'est plus nécessairement vrai que les deux types de sélection sont équivalents. La préférence entre un critère temporel et un critère d'activation sera alors fonction du détail des propriétés des fonctions d'activation.

Les critères de sélection peuvent être rendus plus complexes pour tenir compte d'informations supplémentaires présentes dans les fonctions d'activation. On peut notamment vouloir appréhender l'état d'activation dans le lexique au moment de la sélection en considérant l'ensemble des fonctions d'activation les unes par rapport aux autres. Dans un modèle à critère de sélection basé sur le niveau d'activation, une façon d'arriver à cela est de tenir compte des niveaux d'activation de plusieurs unités dans le critère de sélection : une unité est sélectionnée lorsque son activation dépasse celle de toutes les autres d'une valeur fixée à l'avance (figure 3D ; Wheeldon & Monsell, 1994, discutent cet

³ Dans ces conditions, l'avantage de l'utilisation de critères définis en termes de niveaux d'activation est qu'il permet une interprétation psychologique directe de la dimension non restreinte, le temps (sous réserve bien sûr que les patterns temporels obtenus soient appropriés).

algorithme sans l'implémenter ; voir aussi Starreveld & La Heij, 1996). La même démarche est possible dans un modèle à critère de sélection temporel. (Dell, Burger et al., 1997) utilisent le critère de sélection suivant : au temps n de sélection, toutes les unités ont une chance d'être sélectionnées, et cette chance est proportionnelle à leur niveau d'activation. De cette façon l'unité qui a le plus de chances d'être sélectionnée est bien sûr l'unité la plus active. Mais le système est en mesure de produire des erreurs pour deux raisons : car l'unité la plus active n'est pas nécessairement la cible ou bien car, bien qu'elle soit la plus active, une autre peut être sélectionnée.

Une implémentation relativement similaire à celle-ci est celle utilisée par l'influent et très complet modèle d'accès lexical WEAVER++ (Levelt et al., 1999; Roelofs, 1992, 1997). Dans WEAVER, l'unité à sélectionner n'est pas choisie sur la base de critères d'activation ou de temps. WEAVER implémente des procédures dites de *vérification* qui permettent de déterminer que l'unité sélectionnée à un niveau soit bien celle qui correspond à l'unité sélectionnée à un niveau précédent. Ainsi ce modèle n'est pas en mesure de modéliser les erreurs, puisque ses sélections sont toujours correctes. Par exemple, dans sa modélisation d'une tâche à deux inputs – le paradigme d'interférence mot-dessin – WEAVER sait quelle unité il doit choisir sur la base de drapeaux (*flags*) qui marquent la provenance de l'activation qui arrive dans chaque unité⁴. Toutefois WEAVER prend en compte les activations relatives, en aval de la sélection. Dans ce modèle, la variabilité qui permet de rendre compte des latences de dénomination est basée sur ces activations relatives. Une fois qu'une unité déterminée a été sélectionnée (par le critère externe de marquage), le temps nécessaire à la récupération de ses propriétés est donné par le rapport de son activation par rapport aux activations de toutes les autres unités pertinentes. Autrement dit, contrairement à Dell, Burger et collaborateurs

⁴ Marquer la provenance de l'activation peut paraître un procédé peu naturel. Ceci n'est vrai que si l'on imagine l'activation comme une quantité indiscriminée, échangeable sans contraintes d'une partie à l'autre d'un modèle. Par contre, si on modélise une tâche expérimentale on peut imaginer introduire certains aspects du traitement (par exemple de type attentionnel) qui indiquent l'origine des informations. Une autre façon de modéliser ce type d'hypothèse dans un modèle de tâche serait de définir les séries numériques comme des variables multi-variées, pour lesquelles chacune des sous variables coderait l'activation provenant d'une source différente. Ainsi on pourrait garder une trace de l'activation totale et de l'activation par source.

(1997) qui utilisent les niveaux relatifs d'activation pour générer des erreurs, Roelofs (1997) les utilise pour générer des latences de réponse.

Dans les critères discutés jusqu'ici, la sélection nécessite un critère externe qui détermine l'information sélectionnée : nombre de cycles, seuil absolu ou relatif d'activation, étiquetage des séries numériques etc. Ce critère externe est nécessaire si le modèle livré à lui-même évolue de façon continue sans qu'aucune réponse ne soit définie de façon intrinsèque. Les modèles où les activations des unités convergent définissent d'eux-mêmes les réponses qu'ils atteignent. Par définition, une suite (fonction d'activation) convergente cesse de varier de façon significative au bout d'un nombre de cycles ; en définissant un critère pour cette variabilité⁵ (p.ex. variabilité pas supérieure à 1% de la valeur de la série), on peut modéliser d'une part le temps qui est nécessaire pour atteindre cet état stable, et d'autre part la nature de l'état ainsi atteint –soit une réponse correcte soit une erreur. Ce type de fonction d'activation présente de l'attrait de produire des réponses de façon intrinsèque : un état initial conduit de façon non ambiguë à l'état final correspondant. A ce jour, cette implémentation n'a été utilisée que pour rendre compte de la précision des réponses (ex. Lambon Ralph et al., 2001) mais pas les latences.

Cette discussion des critères de sélection doit aussi mentionner les conséquences de cette sélection : que se produit-il dans le modèle lorsqu'une unité est sélectionnée. En fait, nous avons mentionné ce point dans la discussion de l'interactivité dans les modèles de production (cf. section précédente) : nous avons vu que le degré d'interactivité d'un modèle dépend dans une certaine mesure des conséquences de la sélection à chacun des niveaux. La modélisation de la sélection des états initiaux du système a aussi son importance : ceux-ci peuvent être modélisés par une activation constante des unités représentant l'input, ou bien par une activation décroissante, ou bien par une activation fixée pour un nombre fini de cycles. Cette simple différence conduit à des patterns d'activation pour les unités d'output qui peuvent être très différents (figure 4).

⁵ Comme précédemment, ce critère est externe. Cependant, contrairement aux précédents, il ne sert pas à identifier la réponse produite –qui est déjà connue par la valeur de convergence de la série ; il sert plutôt à fixer un niveau de stabilité de cette réponse.

Comme on peut le constater, un défi dans le domaine de la production est d'élaborer le ou les critères de sélection qui permettront de rendre compte, dans un seul modèle, de patterns de latences et d'erreurs. Un modèle idéal serait celui qui reproduit les patterns de latences de réponses, et qui en même temps tient compte des biais et des régularités observées dans les productions erronées. Mieux encore, le modèle devrait être en mesure de simuler les événements de manque du mot (mot sur le bout de la langue : Brown, 1991; omissions de patients anomiques : Laine et al., 1998), de même pour les autres réponses inhabituelles des patients aphasiques (ex. circonlocutions : Rumel et al., 2000). Raisonner à partir de la définition et l'exigence du critère de sélection utilisé -là où le modèle doit arriver- est sûrement une démarche constructive pour la mise au point d'un modèle complet.

2.3) Lésions de modèles et modélisation de lésions

L'un des objectifs des modèles de production de langage est de simuler les patterns d'erreurs produits par les locuteurs sains ou aphasiques. Dans de nombreux cas, la modélisation d'erreurs se fait en deux étapes. La construction du modèle conduit à formuler des équations et des paramètres produisant essentiellement les réponses correctes (dans certains cas, le modèle du fonctionnement normal aura une capacité marginale à produire des erreurs). Ensuite, les fonctions numériques ou leurs paramètres sont modifiés d'une façon spécifique pour qu'elles ne produisent plus les patterns corrects, mais des patterns déviants attendus. De cette façon la modélisation offre deux « niveaux de liberté » : d'une part dans la définition initiale du modèle et d'autre part dans sa modification ultérieure. La modélisation de la lésion d'un système lexical doit se faire en relation avec l'hypothèse neuropsychologique que l'on cherche à implémenter. On peut par exemple postuler qu'une lésion affectera de façon équivalente l'ensemble des niveaux de traitement impliqués dans une tâche. Cette hypothèse de globalité du dommage a pu être utilisée avec un succès certain dans des modèles de production (Dell, Schwartz et al., 1997; voir discussion dans Dell et al., 2000; Foygel & Dell, 2000), mais elle est restée controversée sur le plan computationnel et neuropsychologique (Rumel & Caramazza, 2000). Les hypothèses alternatives proposent que le dommage peut n'affecter que certains des traitements impliqués dans la réalisation d'une tâche expérimentale, ou que certains niveaux de

traitement sont plus affectés que d'autres par la lésion (*cf.* discussion de la manipulation du paramètre de connexion, ci-dessous).

A la base de la production d'erreurs se trouve dans la très grande majorité des cas la présence de bruit contribuant aux niveaux d'activation des unités. Ce bruit est conçu comme une variable aléatoire, généralement une gaussienne de moyenne nulle dont l'écart type peut dépendre de plusieurs paramètres. Tout d'abord, la sévérité du déficit : l'écart type du bruit sera plus grand pour modéliser un déficit plus grand. Par ailleurs, la magnitude du bruit peut être homogène à travers les unités qu'elle affecte -bruit intrinsèque- où bien être dépendante du degré d'activation de chacune des unités affectées -bruit d'activation. Postuler que le bruit affecte en magnitude les unités de façon homogène peut constituer un problème pour la stabilité du système lexical. Sous cette hypothèse une unité très peu activée peut recevoir la même quantité absolue d'activation (ou d'inhibition) qu'une unité très activée (Rapp & Goldrick, 2000). Dans ces conditions, les patterns d'activation produits dans le modèle standard peuvent subir des bouleversements relativement arbitraires. Par contre, si la magnitude du bruit est liée au niveau d'activation, les déviations de trajectoire des unités dues au bruit préserveront dans une certaine mesure le pattern d'activation présent dans le modèle non lésé. Dans la mesure où l'on postule généralement que les erreurs sont contraintes par les structures du fonctionnement normal (Fromkin, 1971), ce deuxième choix de bruit semble plus approprié pour l'accès lexical en production.

En plus de manipuler la magnitude et la nature du bruit affectant l'activation des unités, les paramètres définissant le fonctionnement normal des modèles peuvent être modifiés de différentes façons pour simuler diverses lésions. Nous passons à présent en revue quelques unes des caractéristiques de ces manipulations qui ont été évaluées, par exemple, à l'aide de simulations (Dell, Schwartz et al., 1997; Foygel & Dell, 2000; Rumel et al., 2000).

Le premier paramètre que l'on peut manipuler pour simuler une lésion est le paramètre d'atténuation de l'activation (d). D'après l'équation générale présentée en p. ###, ce paramètre module la quantité d'activation propre qu'une unité maintient d'un essai sur l'autre. Plus d est élevé (proche de 1), plus l'activation de l'unité ressemblera à celle qu'elle avait au cycle précédent. De façon très approximative on peut donc dire que ce paramètre indexe une mesure de la stabilité du système. La

modulation de d est l'une des manipulations utilisée par Dell, Schwartz et collaborateurs (1997) et par Foygel et Dell (2000) pour créer l'espace de modèles possibles où ils proposent de localiser les patients aphasiques dont ils essayent de simuler la performance. Leur étude de simulations montre que la modulation de d est en partie responsable des variations dans les proportions d'erreurs de type sémantique et phonologique (par opposition aux erreurs non-mots qui n'ont pas de relation avec la cible attendue). Ils attribuent cette observation au fait que la manipulation de d par elle-même maintient intacte la structure informationnelle du réseau. Donc les opportunités d'erreurs sont modelées par la proximité dans le réseau (d'où des erreurs reliées). De fait Rapp et Goldrick (2000) argumentent que la manipulation de d n'a d'autre effet que réduire le niveau d'activation global présent dans le système et donc de rendre visibles les effets du bruit.

Un point de vue légèrement différent est adopté par Rumel et collaborateurs (2000). Leur manipulation du paramètre d les conduit à un ensemble très restreint de modèles possibles. En d'autres termes : ce paramètre module très faiblement le fonctionnement du modèle qu'ils examinent et cette manipulation ne produit pas des patterns d'erreurs très variés. En conséquence, ils questionnent l'utilité de la manipulation de ce paramètre. En outre, une analyse de l'ensemble des valeurs prises par d dans les ajustements des modèles de Dell, Schwartz et collaborateurs (1997) ou Rumel et collaborateurs (2000) montre que la variabilité de d est relativement plus faible que la variabilité du paramètre p de connexion entre les unités. Cela suggère que le paramètre d pourrait jouer un rôle moins important que le paramètre de connexion dans l'ajustement du modèle aux performances des patients.

Une raison possible du rôle plus ou moins important attribué à ce paramètre pourrait être le contexte dans lequel cette manipulation est faite. Comme nous l'avons noté précédemment, l'équation de propagation comporte plusieurs termes. L'importance de chacun de ceux-ci pour le fonctionnement global du modèle dépend de leurs valeurs relatives. Dans un modèle où la valeur de $p \times m$ (paramètre moyen de connexion fois nombre moyen de connexions) est importante par rapport à d ce dernier paramètre devrait avoir un rôle plus réduit dans la mise à jour des activations des unités. Ainsi, bien que la manipulation exclusive de d nous renseigne sur le rôle de ce paramètre dans le modèle, elle ne peut se substituer à une analyse combinée des différentes contributions à la propagation d'activation.

Un autre type de paramètre qui a pu être manipulé pour simuler des lésions dans un modèle d'accès lexical est le paramètre p caractérisant la transmission d'information entre les unités (connexions). On conçoit généralement la connectivité dans le modèle comme l'élément essentiel du codage d'information dans le système. Ainsi une manipulation de ces connexions va directement affecter l'information qui est présente dans le système. Une conséquence directe de cela est que la manipulation de la connectivité peut donner lieu à de nombreux résultats différents, selon comment l'information codée est affectée. On peut noter que l'effet de ces manipulations dépendra étroitement de la façon dont la propagation d'information est implémentée dans le modèle étudié. Dans certains modèles d'accès lexical, toutes les connexions ont le même poids (l'information est alors codée dans le système par la présence ou absence de lien entre des unités). Sous cette hypothèse, on peut baisser la valeur du paramètre de transmission de façon globale, ce qui affectera l'ensemble du fonctionnement du système. Cela revient à limiter le poids de l'information codée dans les connexions dans le processus de propagation d'activation. Dell, Schwartz et collaborateurs (1997) montrent que cette manipulation conduit à une baisse de la capacité du modèle à « séparer » les niveaux d'activation des différentes unités, ce qui se traduit par une augmentation des erreurs aléatoires (interprétées comme réponses non-mots n'ayant pas de relation claire avec la réponse attendue). Une variante de cette manipulation consiste à catégoriser les connexions, et de n'autoriser la manipulation que d'une partie d'entre elles : par exemple celles qui relient deux niveaux spécifiques (ex. modèle sémantique phonologique de Foygel & Dell, 2000). Cette hypothèse de lésion est bien sûr beaucoup plus structurée que la précédente, dans la mesure où elle intègre des distinctions basées sur la théorie et les observations neuropsychologiques (ex. distinction sémantique vs. phonologie). De façon non surprenante, elle permet de rendre compte de déficits beaucoup plus spécifiques, comme par exemple la production d'une forte proportion d'erreurs d'un seul type (voir aussi Rapp & Goldrick, 2000; Rumel et al., 2000).

Dans certains modèles, notamment ceux qui utilisent des représentations distribuées, les connexions ne sont pas toutes identiques. Dans ce genre de modèles, postulant généralement une très grande connectivité, le processus d'apprentissage conduit à un codage d'information basé sur la modulation du poids des connexions entre unités. Dans ce cas là les lésions sont généralement

modélisées en annulant une proportion des connexions. Par exemple on peut remettre à zéro 25 % ou 50% des connexions (choisies aléatoirement) entre deux niveaux de traitement (Lambon Ralph et al., 2001). L'impact de ce type de lésion est encore plus difficile à évaluer *a priori* en raison du caractère distribué de la représentation d'information. Dans les faits, seule une étude détaillée des conditions de la lésion, particulièrement si elles tiennent compte des autres paramètres inclus dans le modèle pourra permettre de caractériser l'impact d'une lésion des connexions sur le fonctionnement du modèle. A ce jour ce genre d'étude n'a été envisagé que sous forme numérique, et une étude sous forme exacte serait peut être souhaitable.

Pour finir cette section, nous pouvons citer une méthode relativement différente et peu utilisée de modéliser la production d'erreurs. Cette méthode consiste à postuler que le modèle bruité produit des erreurs avec une probabilité p_E spécifiée. Au cours d'un essai expérimental, on détermine par tirage de probabilité p_E si l'essai sera bruité ou pas. S'il ne l'est pas, le fonctionnement normal du modèle est respecté. Si l'essai est bruité, alors on ne laisse pas le modèle sélectionner la réponse qu'il aurait naturellement choisie mais on le force à en sélectionner une autre parmi les restantes, sur la base de leurs niveaux d'activation. Vousden et collaborateurs (2000) implémentent ce système, en rajoutant une deuxième contrainte : l'item choisi « par erreur » ne doit pas être plus distant qu'un critère fixe de l'item cible. On peut noter que cette modélisation de l'apparition des erreurs reste en dehors du processus décrit. La définition de production d'erreurs ainsi obtenue ne constitue donc pas une hypothèse sur le mécanisme qui produit les erreurs, ce qui peut être vu comme une faiblesse de cette approche par rapport aux approches décrites précédemment.

3) Conclusion

Dans ce chapitre nous avons passé en revue un certain nombre des implémentations computationnelles d'hypothèses psychologiques utilisées pour rendre compte des processus de production de langage. Ce bref panorama a permis de mettre en évidence la diversité des options existant pour rendre compte des mêmes hypothèses ou d'hypothèses très proches. Cette diversité complique la tâche de comparaison entre les modèles ainsi que le choix du modèle « le plus pertinent » ou « le plus vrai ». La diversité actuelle devrait pourtant permettre à terme de converger sur les

caractéristiques computationnelles les plus appropriées pour obtenir une modélisation exhaustive du processus d'accès lexical. De fait les modèles existant peuvent être vus de deux façons : soit comme des descriptions les plus fidèles possibles aux traitements mis en jeu dans le comportement, soit comme des outils permettant d'explorer de façon systématique *différentes descriptions possibles* du comportement étudié. A ce jour, cette deuxième option apparaît comme le passage obligé pour atteindre la première.

Légendes des figures

Figure 1 :

Gauche : Représentation schématique du mini modèle de production formulé à titre illustratif. Ce mini modèle partage avec de nombreux modèles de production de parole une structure en trois niveaux, sans connexions inter niveaux et sans connexions du premier au troisième niveaux (A → C). Droite : Exemple de niveaux d'activation des unités de sortie (paramètres de transmission d'activation fixés aléatoirement).

Figure 2 :

Codage de l'entrée lexicale du mot « différence » dans un modèle localiste (le codage est assuré par l'activation de l'une des unités, B-13) et dans un modèle distribué (le codage est constitué par un vecteur de valeurs d'activation caractérisant l'ensemble des unités).

Figure 3 :

A- Critère de sélection temporel : l'unité sélectionnée (C4) est celle qui est le plus activée au temps $n = 10$

B- Critère de sélection d'activation : l'unité sélectionnée (C4) est celle qui la première dépasse un seuil d'activation fixé à l'avance.

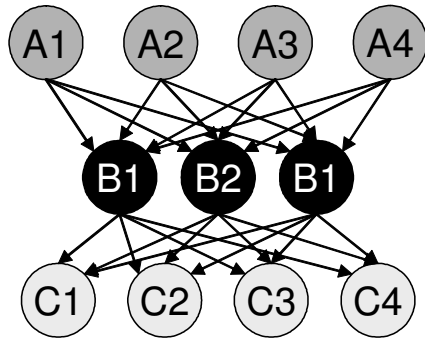
C- Fonctions d'activation ne permettant pas d'établir une équivalence entre les critères de sélection temporel et d'activation ; dans ce type de situation, le choix du critère de sélection nécessitera la prise en compte détaillée des fonctions d'activation.

D- Critère de sélection basé sur la différence d'activation : l'unité sélectionnée (C4) est celle qui la première dépasse les autres d'une activation Δa_i fixée à l'avance.

Figure 4 :

Courbes d'activation des séries d'output dans un modèle où l'activation de l'input est maintenue pendant toute la durée du traitement (C1), dans un modèle où l'activation de l'input est constante pendant une durée fixe (C2), et dans un modèle où l'activation de l'input est décroissante pendant une durée fixe (C3).

Figure 1



$$a_i(t) = 0.5 \quad [1 < t < 10]$$

$$u_i(t) = u_i(t-1) * 0.6 + \sum [u_j(t-1) * p_{ij}]$$

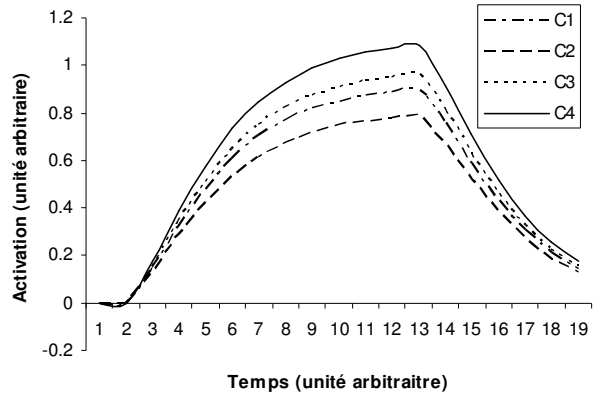


Figure 2

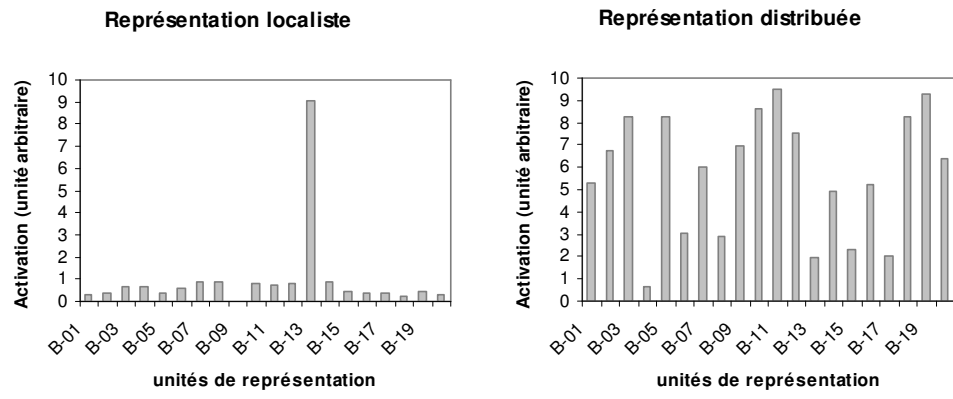


Figure 3

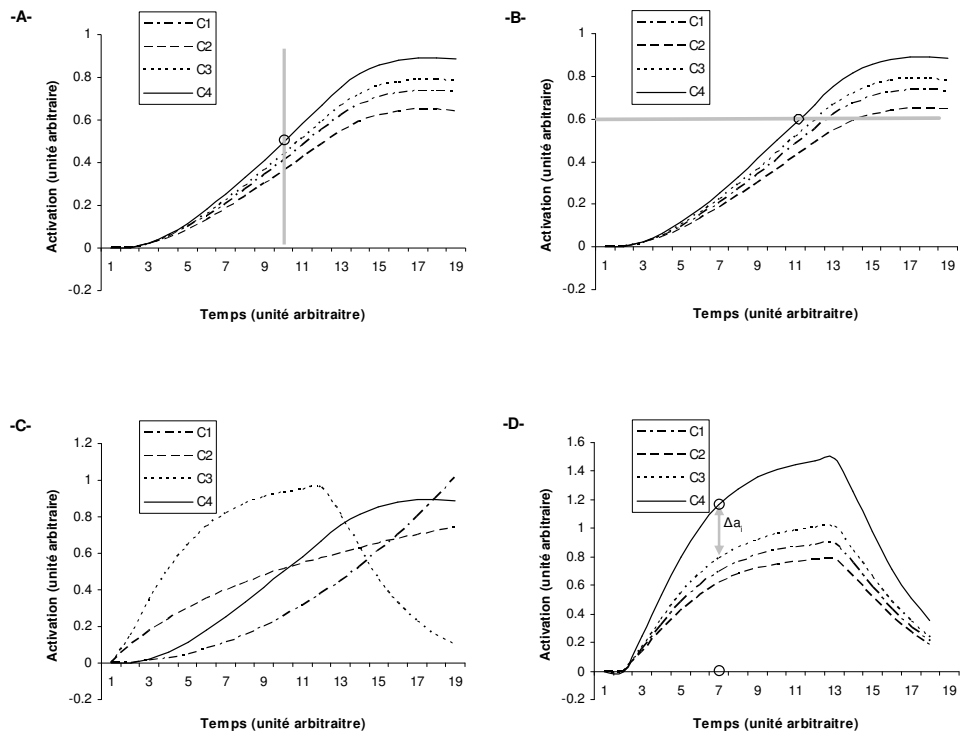
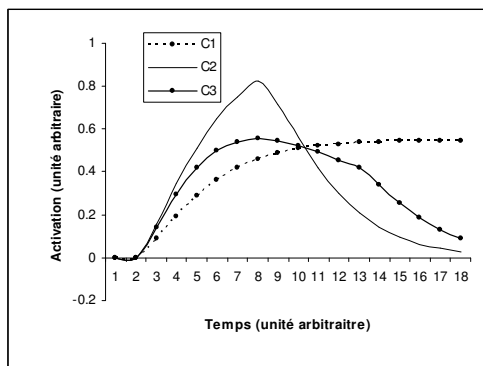


Figure 4



REFERENCES

- Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review*, *105*(4), 678-723.
- Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*, 204-223.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.
- Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283-321.
- Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production in serial order: A functional analysis and a model. *Psychological Review*, *104*, 123-144.
- Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production. *Cognitive Science*, *17*, 149-195.
- Dell, G. S., Martin, N. M., Saffran, E. M., Schwartz, M. F., & Gagnon, D. A. (2000). The role of computational models in neuropsychological investigations of language: Reply to Ruml and Caramazza (2000). *Psychological Review*, *107*(3), 635-645.
- Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al (1991). *Psychological Review*, *98*, 604-614.
- Dell, G. S., & O'Seaghdha, P. G. (1994). Inhibition in interactive activation models of linguistic selection and sequencing. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 409-453). San Diego, CA: Academic Press, Inc.
- Dell, G. S., Schwartz, M. F., Martin, N. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801-838.
- Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory & Language. Special Issue: Disorders of language and memory: Implications for cognitive theory*, *43*(2), 182-216.
- Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*, 27-52.
- Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, *8*, 291-309.
- Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In the PDP research group (Ed.), *Parallel distributed processing*. Cambridge, MA: The MIT Press.
- Humphreys, G. W., & Riddoch, M. J. (1988). Cascade processes in picture namig. *Cognitive Neuropsychology*, *5*, 67-104.
- Laine, M., Tikkala, A., & Juhola, M. (1998). Modelling anomia by the discrete two-stage word production architecture. *Journal of Neurolinguistics*, *11*(3), 275-294.
- Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No right to speak? The relationship between object naming and semantic impairment:

- Neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience*, 13(3), 341-356.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1-75.
- MacKay, D. G. (1987). *The organization of perception and action*. New York, NY: Springer Verlag.
- Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral & Brain Sciences*, 23(4), 443-512.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103(1), 56-115.
- Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10(5), 377-500.
- Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, 107(3), 460-499.
- Roelofs, A. (1992). A spreading activation theory of lemma retrieval in speaking. *Cognition*, 42, 107-142.
- Roelofs, A. (1997). The WEAVER model of word-form encoding in spoken production. *Cognition*, 64, 249-284.
- Rumel, W., & Caramazza, A. (2000). An evaluation of a computational model of lexical access: Comment on Dell et al. (1997). *Psychological Review*, 107(3), 609-634.
- Rumel, W., Caramazza, A., Shelton, J. R., & Chialant, D. (2000). Testing assumptions in computational theories of aphasia. *Journal of Memory & Language. Special Issue: Disorders of language and memory: Implications for cognitive theory*, 43(2), 217-248.
- Santiago, J., MacKay, D. G., Palma, A., & Rho, C. (2000). Sequential activation processes in producing words and syllables: Evidence from picture naming. *Language & Cognitive Processes*, 15(1), 1-44.
- Starreveld, P. A., & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 896-918.
- Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, 41(2), 101-175.
- Wheeldon, L. (2000). Aspects of language production.
- Wheeldon, L. R., & Monsell, S. (1994). Inhibition of spoken word production by priming a semantic competitor. *Journal of Memory and Language*, 33, 332-356.
- Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception & Performance*, 24(4), 1131-1161.

Note

L'auteur remercie Johannes Ziegler et Stéphane Dufau pour leurs commentaires sur ce manuscrit.