Modélisation et simulation par carte auto-organisatrice de l'effet de fréquence des mots chez l'apprenti lecteur

S.Dufau[1], B. Lété[2], C. Touzet[3], H. Glotin[4], J. Ziegler[1] & J. Grainger[1]
[1] Laboratoire de Psychologie Cognitive, UMR6146, CNRS & Université de Provence
[2] INRP & Laboratoire d'Etude des Mécanismes Cognitifs, UMR5596, CNRS & Université de Lyon 2
[3] Laboratoire de Neurobiologie Intégrative et Adaptative, UMR6149, CNRS & Université de Provence
[4] LSIS, UMR6168, CNRS & Université du Sud, Toulon Var

Stéphane Dufau, Laboratoire de Psychologie Cognitive, Centre Saint Charles - Case D, Université de Provence, 3 place Victor Hugo, 13331 MARSEILLE cedex 3, France. stephane.dufau@univ-provence.fr

RESUME
Un réseau de neurones artificiels de type carte auto-organisatrice a été utilisé pour simuler l'apprentissage des formes orthographiques des mots du français chez l'apprenti lecteur. Entraîné à partir des mots vus par les enfants, ce modèle génère des résultats similaires à ceux issus d'une expérience comportementale de décision lexicale. Cette expérience a mesuré les performances des enfants de langue française âgés de 6 à 11 ans et met en évidence un effet de fréquence lexicale (les mots de haute fréquence sont reconnus plus aisément que ceux de basse fréquence) que le modèle est capable d'intégrer.

MOTS CLES
Carte auto-organisatrice ; Apprentissage ; Lecture ; Apprentissage de la lecture.

1. Introduction

L'effet de fréquence lexicale est certainement le phénomène le plus répliqué dans les études psycholinguistiques. Il reflète le fait que les mots apparaissant plus fréquemment dans une langue donnée sont perçus plus rapidement et produisent moins d'erreur de reconnaissance que des mots apparaissant moins fréquemment. Cet effet est consistant aussi bien à travers les tâches expérimentales (décision lexicale, dénomination, identification perceptive, catégorisation sémantique) qu'à travers les langues [1].

Depuis le modèle logogène de Morton (1969), de nombreuses tentatives de modélisation de reconnaissance des mots écrits ont été effectuées. Ainsi, la première génération de modèles neuro-computationnels (modèle d'activation interactive, modèle de recherche en série et modèle d'activation-vérification) utilise un paramètre dont la valeur, choisie *a priori* par le modélisateur selon la classe d'appartenance des mots, permet de rendre compte de l'effet de fréquence. Par exemple, un seuil fixé plus haut pour les mots de basse fréquence que pour ceux de haute fréquence. Un mot de haute fréquence passe ainsi plus rapidement et plus facilement son seuil de reconnaissance. Ces modèles n'expliquent pas des éléments fondamentaux comme l'effet de fréquence émergeant de l'exposition aux mots parlés et/ou écrits. De plus, ces modèles sont *ad-hoc* au sens où la valeur des poids de connexions est fixée par l'expérimentateur. La seconde génération de modèles a pris en compte ces critiques et utilise des réseaux de neurones artificiels dotés de capacité d'apprentissage [2]. Si le seuil de reconnaissance est toujours fixé *a priori* par l'expérimentateur, il est devenu identique quelle que soit la classe des mots. La variabilité nécessaire est ici induite par l'ajustement des poids de connexions entre la représentation orthographique des mots et leur représentation lexicale. Cette variabilité découle de la fréquence et de l'ordre d'apparition des mots dans la base d'apprentissage.

Le travail présenté ici s'inscrit dans une volonté de proposer une modélisation plus pertinente. Nous avons utilisé un réseau de neurones artificiels de type carte auto-organisatrice (Kohonen [3]). Ce réseau a été entraîné à partir des mots de la base lexicale francophone Manulex [4] qui répertorie les fréquences des mots d'un ensemble de manuels scolaires par groupe d'âge. Nos simulations nous ont permis d'obtenir des résultats dont la variabilité se rapproche de celle des résultats comportementaux de reconnaissance des mots écrits, recueillis auprès d'enfants de 6 à 11 ans (répartis en 5 groupes de niveau CP à CM2) scolarisés en France. L'expérience consistait en une tâche de décision lexicale manipulant la fréquence des mots (conditions haute et basse) et le voisinage orthographique (conditions haute et basse) calculés à partir de la base lexicale Manulex.

Nous présentons dans une première partie les résultats de l'étude comportementale. Dans une seconde partie, nous présentons notre modèle et les simulations effectuées dans le même contexte que l'étude précédente. Enfin, nous concluons sur la validité de notre modélisation et les perspectives explicatives qui en découlent.

2. Expérience comportementales

20 participants ont été sélectionnés pour chaque niveau d'études allant du CP au CM2 (excepté en CM1, 10 participants) sur un test d'habileté en lecture. 56 mots

de 4 et 5 lettres ont été choisi dans la base lexicale Manulex répondant aux critères d'équilibre croisé, pour chaque groupe d'age, de haute et basse fréquence lexicale (28 mots HF et 28 mots BF) ainsi qu'aux critères de haut et bas voisinage orthographique (28 mots HV et 28 mots BV), donnant ainsi 14 mots par condition. 56 nonmots (28 prononçables et 28 non prononçables) ont également été générés, ce qui porte à 112 stimuli visuels vus par chaque participant. L'expérience consistait en une tâche de décision lexicale effectuée sur ordinateur. Des mesures de temps d'identification (temps de réaction TR entre l'apparition d'un stimulus à l'écran et l'appui d'une touche sur le clavier) et de bonne reconnaissance (l'apparition d'un mot déclenche une réponse positive) ont été effectuées.

Les mesures du taux de bonne reconnaissance sont rapportées au tableau 1. Des analyses statistiques de type ANOVA ont été effectuées sur ces données. Leurs résultats sont résumés dans le tableau 2.

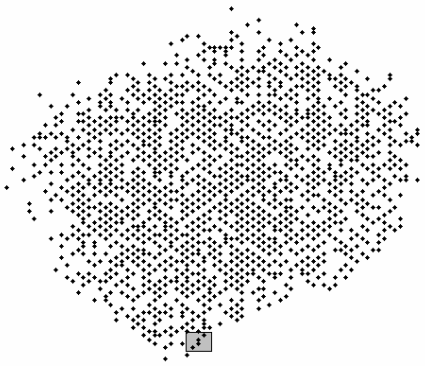
Tableau 1. Pourcentage de bonne réponse par niveau d'études en fonction des conditions de fréquence lexicale (haute et basse fréquence ; HF et BF) et de voisinage orthographique (haut et bas voisinage ; HV et BV). D'après les travaux de l'ACI-CNRS [5].

	CP	**CE1**	**CE2**	**CM1**	**CM2**
HF	82	92	96,1	95,7	98,2
BF	50,9	67,1	76,9	82	84,5
HV	69,8	81,2	87,8	88,9	91,3
BV	63	77,9	85,2	88,8	91,5

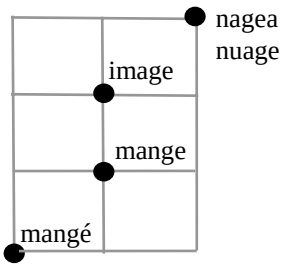
L'étude examine l'effet de fréquence et de voisinage orthographique à travers 5 niveaux d'études (CP au CM2, soit de 6 à 10 ans environ). Le taux de bonne reconnaissance a été la seule variable dépendante utilisée, le temps de reconnaissance variant de plusieurs centaines de millisecondes chez les enfants les plus âgés à plusieurs secondes chez les plus jeunes. Comme le montre le tableau 1, les mots de haute fréquence sont reconnus plus aisément que ceux de basse fréquence (par exemple au CP, 82% contre 50,9%). L'effet du voisinage orthographique est moins visible. Selon les résultats statistiques présentés au tableau 2, les effets simples de fréquence et de niveau sont présents sur toutes les analyses, ainsi que l'interaction niveau x fréquence. L'effet de fréquence obtenu corrobore les résultats de Ducrot *et al* [6]. qui l'avait étudié chez les enfants français de CP. Les effets niveau et fréquence corroborent également les résultats de Burani *et al.* [7] chez les enfants italiens de 9 à 11 ans (l'expérience n'ayant pas conduit à une interaction entre ces deux facteurs). Il est à noter enfin l'effet simple de voisinage, suggérant le rôle de l'identification des mots par une stratégie d'analogie chez les enfants les plus jeunes.

3. Modèle et simulation

Une carte auto-organisatrice incrémentale a été entraînée sur une base d'apprentissage comportant environ 2500 mots de 4 et 5 lettres. Nous avons utilisé 10 itérations d'apprentissage par niveau d'études et nous avons réalisé 24 apprentissages (soit 24x10 itérations d'apprentissage par niveau d'études). L'objectif est de permettre des tests ANOVA sur les performances en apprentissage du réseau de neurones. Chaque mot a été codé au préalable par un vecteur unique de 1681 dimensions représentant un sous-ensemble des couples de lettres possibles du mot appelés bigrammes ouverts [8]. Par exemple, le mot TABLE est codé par les bigrammes (TA ; TB ; TL ; AB ; AL ; AE ; BL ; BE ; LE). Ces formes orthographiques sont projetées au cours de l'apprentissage sur une carte auto-organisatrice ; chaque forme orthographique y est représentée par un neurone. Si un neurone est associé à une forme orthographique unique, ce neurone est considéré comme fournissant une réponse correcte. A l'inverse, un neurone associé à plusieurs formes orthographiques est considéré comme fournissant une réponse incorrecte. Ainsi, l'unicité de la représentation lexicale d'une forme orthographique permet de simuler l'état de la connaissance lexicale d'un enfant. Par exemple, les neurones associés à une forme orthographique à l'issu d'un processus d'apprentissage sont représentés sur la figure 1a. La portion de la carte auto-organisatrice (figure 1b) montre des mots groupés suivant leur similarité orthographique. On remarque également que les mots fréquents (image, mange) sont représentés par un neurone chacun alors que les mots peu fréquents (nagea, nuage) partage un même neurone.



(a)



(b)

Figure 1. (a) Carte auto-organisatrice après apprentissage (80 x 80 neurones). Les points représentés en noir sont les neurones associés à une ou plusieurs formes orthographiques ; les neurones non associés à un mot ne sont pas représentés; en gris, le détail de la figure 1b ; (b) une portion de la carte auto-organisatrice de 3 x 4 neurones.

4. Résultats

La figure 2 montre le pourcentage de réponses correctes en fonction du niveau d'études, le grade (CP : 1, CE1 : 2, CE2 : 3, CM1 : 4, CM2 : 5), et du facteur fréquence où HF et LF sont les conditions de haute et basse fréquence. La figure 3 montre ce pourcentage concernant le facteur voisinage où HN et LN sont les conditions de haut et bas voisinage orthographique. L'écart type rapporté sur les figures 2 et 3 (Root Mean Square Deviation : RMSD) mesure l'écart par condition entre les performances des enfants et celle du modèle.

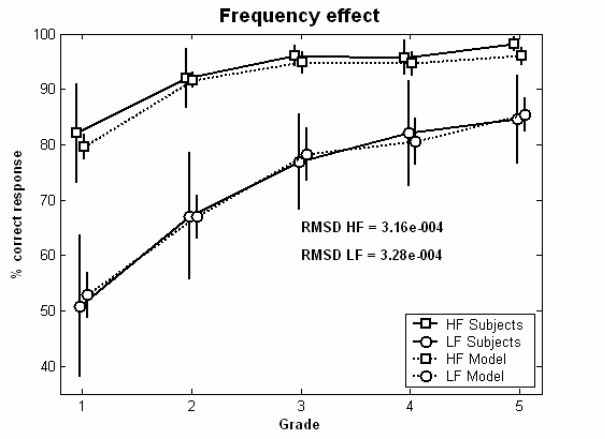


Figure 2. L'effet de fréquence suivant le niveau d'études. Les résultats des enfants sont représentés en traits pleins et ceux du modèle en pointillés.

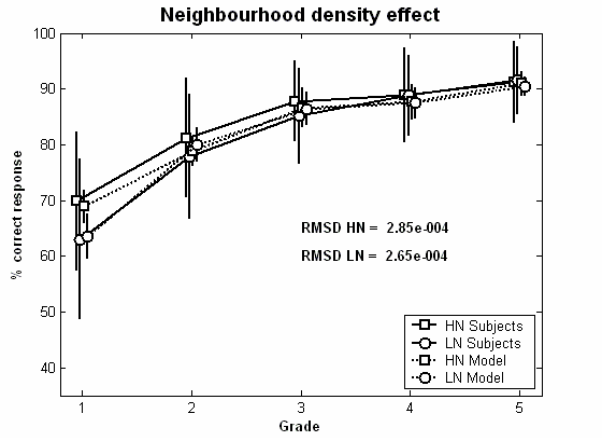


Figure 3. L'effet de voisinage suivant le niveau d'études. Les résultats des enfants sont représentés en traits pleins et ceux du modèle en pointillés.

Une analyse de variance identique à celle effectuée sur les données comportementales a été réalisée. Les résultats figurent dans le tableau 2. On remarque l'adéquation des résultats en F2 entre ceux des enfants et ceux des simulations avec des effets simples du niveau d'études et de la fréquence, ainsi qu'une interaction des facteurs niveau d'études et fréquence. Cette adéquation est également traduite dans les valeurs similaires de l'écart type par condition.

Tableau 2. Valeurs de l'analyse de variance pour les enfants et les simulations (F1) et pour les mots (F2). La significativité est donnée par (*), (**) et (***), respectivement $p < 0.05$, $p < 0.01$ et $p < 0.001$. N est le facteur Niveau, V le Voisinage et F la Fréquence.

	Enfants		**Simulations**	
	F1	**F2**	**F1**	**F2**
N	$25,9^{***}$	$54,8^{***}$	$105,7^{***}$	$40,3^{***}$
F	$215,1^{***}$	$31,70^{***}$	$533,3^{***}$	$45,4^{***}$
V	$7,4^{**}$	0,54	2,1	0,5
NxF	$6,49^{***}$	$7,8^{***}$	$14,9^{***}$	$2,8^{*}$
NxV	2,1	1,3	$2,7^{*}$	0,02
FxV	0,7	0,04	$4,9^{*}$	0,5
NxFxV	0,7	0,4	$10,8^{***}$	2,1

5. Discussion

Une carte auto-organisatrice a été utilisée pour modéliser l'apprentissage de la forme orthographique des mots à travers 5 niveaux d'études. Les performances du modèle sont globalement proches de celles des enfants, notamment l'analyse en F2 qui généralise le comportement du modèle à l'ensemble des mots de 4 à 6 lettres du français. Les valeurs similaires de RMSD selon les conditions indiquent la capacité du modèle à simuler l'influence des facteurs fréquence et voisinage.

L'objet de cette étude a été de développer une carte auto-organisatrice qui reflète l'effet de fréquence et de voisinage orthographique avec le niveau scolaire. Si l'on peut penser que l'effet de fréquence s'installe et augmente à mesure que les enfants apprennent à lire, les données expérimentales montrent l'inverse : diminution avec l'âge. Ce résultat s'explique à partir de la distribution des mots dans les manuels scolaires. Les manuels de CP sont en effet constitués quasi uniquement de mots de haute fréquence, les mots de basse fréquence étant introduits graduellement dans les manuels ultérieurs jusqu'à refléter pour les manuels de CM2 la distribution des mots dans les livres pour lecteurs aguerris.

6. Conclusion

Notre étude a montré qu'un modèle de carte auto-organisatrice pouvait reproduire les effets de fréquence lors de l'acquisition de la lecture. Nos résultats renforcent l'hypothèse selon laquelle la construction du lexique orthographique chez l'enfant est un processus guidé uniquement par les données auxquelles ils sont exposés, pour peu que ces données soient encodées sous forme de bigrammes ouverts.

Notre étude suggère également que les propriétés de conservation des densités de probabilité d'apparition

des exemples de la base d'apprentissage ainsi que de la topologie de l'espace d'entrée (ici, l'espace bigrammes), à la base du fonctionnement de la carte auto-organisatrice, sont à même de proposer une hypothèse sur l'organisation corticale des mots lus chez l'apprenti-lecteur.

L'hypothèse d'un processus auto-organisé d'apprentissage des mots et la description de l'organisation corticale du lexique mental sont à même de faire évoluer notre compréhension des mécanismes cognitifs de la lecture. Notons cependant que notre modélisation ne rend pas compte actuellement des données électrophysiologiques et anatomiques qui montrent que la zone impliquée dans la reconnaissance des mots écrits est moins étendue à l'age adulte que chez les enfants, et que cette zone migre pendant l'apprentissage. Notre modélisation pourrait s'étendre également à la description des processus cognitifs d'identification des mots écrits en intégrant notamment les notions du temps de réaction et de voie phonologique.

La connaissance des mécanismes cognitifs de construction et d'utilisation du lexique mental chez l'enfant est certainement un atout pour améliorer l'enseignement d'une langue et parvenir à sa maîtrise. Elle permet d'évaluer les méthodes d'apprentissage de la lecture (globale, semi-globale, phonétique), de mesurer l'efficacité des manuels de lecture ou de créer le manuel « idéal ». Le modèle proposé dans cet article, biologiquement plausible, pourrait également servir de modèle expérimental pour l'étude des troubles de la lecture comme la dyslexie (qui touche 8% des élèves de l'école primaire), modèle sur lequel on évaluerait l'efficacité des méthodes de remédiation.

References

[1] S. Monsell, *The nature and locus of word frequency effects in reading. Basic processes in reading: Visual word recognition*. (D. Besner and G. W. Humphreys. Hillsdale, NJ, England, Lawrence Erlbaum Associates 350 pp: 148-197, 1991).

[2] J. Grainger, and T. Dijkstra. *Visual word recognition: Models and experiments. Computational psycholinguistics: AI and connectionist models of human language processing.* (T. Dijkstra and K. de Smedt. Philadelphia, PA, Taylor & Francis: 139-165, 1996).

[3] T. Kohonen, *Self-Organizing Maps*. (Springer-Verlag, Berlin-Heidelberg-New York, 1995).

[4] B. Lete, L. Sprenger-Charolles, et al., MANULEX: A grade-level lexical database from French elementary school readers, *Behavior Research Methods, Instruments and Computers 36*(1), 2004, 156-166.

[5] J. Grainger, H. Glotin, B. Lété, C. Touzet, J.C. Ziegler & S. Dufau. Modélisation computationnelle de l'apprentissage des mots écrits. Rapport de fin de projet TCAN (ACI-CNRS, 2003-2005).

[6] S. Ducrot, B. Lete, et al., The Optimal Viewing Position Effect in Beginning and Dyslexic Readers, *Current Psychology Letters: Behaviour, Brain and Cognition, 10*(1), 2003.

[7] C. Burani, S. Marcolini, et al., How early does morpholexical reading develop in readers of a shallow orthography?, *Brain and Language 81*(1-3), 2002, 568-586.

[8] J. Grainger and W. Van Heuven, *Modeling Letter Position Coding in Printed Word Perception. Mental lexicon: "Some words to talk about words"* (Nova Science Publishers, Inc., pp: 1-23, 2003).

Remerciements

Ce projet a fait l'objet d'un soutien du programme interdisciplinaire du CNRS ACI-TCAN (2003-2005).