"ADAPTIVE RESONANCE THEORY" : L'ORDRE DE PRÉSENTATION DES MOTS COMPTE-T-IL ?

P. Warnier[1], H. Glotin[1], F. Dandurand[2], S. Dufau[4], S. Fraihat[1], C. Touzet[2], B. Lété[3], J. Ziegler[4] & J. Grainger[4]
[1] Laboratoire des Sciences de l'Information et des Systèmes, UMR6168, CNRS & Université du Sud, Toulon VAR
[2] Laboratoire de Neurobiologie Intégrative et Adaptative, UMR6149, CNRS & Université de Provence
[3] INRP & Laboratoire d'Etude des Mécanismes Cognitifs, UMR5596, CNRS & Université de Lyon 2
[4] Laboratoire de Psychologie Cognitive, UMR6146, CNRS & Université de Provence
Pierre Warnier, pierre@warnier.net & Hervé Glotin, glotin@univ-tln.fr

RESUME
La modélisation de l'apprentissage de la lecture est une nouvelle problèmatique. Dans Neurocomp06, nous avions pu modéliser les performances naturelles de l'enfant en ne donnant que la fréquence des mots dans les corpus à une carte auto-organisatrice (SOM), les mots étant codés par bigrammes ouverts. Cependant SOM ne pemettait pas l'analyse de l'infuence de l'ordre d'apparition des mots dans les corpus. Pour répondre à cette question, nous construisons dans ce papier des modèles neurocomputationnels de type Adaptive Resonance Theory, sur différents niveaux de vigilance, et sur les corpus réels qui contiennent donc plus d'informations qu'une simple fréquence de mots. Nous montrons premièrement que les facteurs fréquence et vigilance sont les plus influents sur les performances. Deuxièment, en présentant des corpus mélangés, nous montrons que l'ordre de présentation des mots est plus influent sur l'apprentissage des mots fréquents, pour les modèles les plus vigilants. Nous proposons finalement une extension de ce nouveau protocole qui amènerait à discuter la construction de corpus.

MOTS CLES
ART2 ; Réseaux de neurones ; Apprentissage ; Lecture ; Vigilance

1 Introduction

Si l'effet de fréquence lexicale est certainement le plus répliqué dans les modèles d'apprentissage de la lecture ([2, 3, 4, 5, 6, 7, 8, 9, 10]), l'ordre d'apparition des mots dans les séquences à été très peu étudié. La Théorie de Résonnance Adaptive (ART2) nous permet ici d'aborder cette question. ART, puis ART2, furent développées dans le but d'éviter le dilemne stabilité-plasticité dans les réseaux d'apprentissage compétitif. L'enjeu est de préserver le savoir précédemment acquis tout en restant capable d'apprendre de nouvelles catégories. Or ceux des architectures ART2 peuvent s'auto-organiser pour produire une reconnaissance stable. Le travail présenté ici fait suite à l'étude duale sur l'effet de fréquence avec un modèle de carte auto-organisée (Kohonen [6][7]). Nous reprenons le même protocole mais en présentant l'intégralité des mots des corpus. Le modèle adéquat pour cette exigence est le (ART2). Cela nous permet de nous concentrer sur l'influence de l'ordre des mots présentés durant l'apprentissage.

Les corpus de mots utilisés pour les expériences sont les livres d'étude "Arthur" (CE1 : 25514 mots, CE2 : 31197, CM1 : 28487, CM2 : 34947) et "Gafi" (CP : 17790 mots) utilisés en classes primaires par les enfants, correspondants aux 5 niveaux du CP au CM2 [6,11]. Nous avons codé la forme graphique de chaque mot comme dans [6,7] avec des bigrammes ouverts des lettres du mots (il en existe 1681 en français), pondérés selon leur position dans le mot ce qui simule des propriétés visuelles [6]. Dans nos expériences préliminaires nous avons montré que cette pondération donne de meilleurs résultats qu'un simple codage binaire des bigrammes. Ainsi le vecteur codant le mot TABLE est le vecteur de 1681 dimensions, nul partout sauf aux indices des bigrammes TA, TB, AB, AL, BL, BE, LE.

2 L'algorithme ART2

ART2 ([1]) est un algorithme d'apprentissage non supervisé : il est capable d'apprendre à reconnaître un vecteur qu'on lui présente en fonction des catégories qu'il construit. Il auto-organise les catégories et en crée quand cela lui semble nécessaire. Le seul contrôle que l'on possède sur ce système d'apprentissage est son paramètrage :

- ***n*** La taille des données à traiter ($0 < i \leqslant n$).
- ***m*** Le nombre de catégories ($0 < j \leqslant m$).
- ***a, b*** Les poids fixes des connexions internes à la couche F_1.
- ***e*** Evite les divisions par 0. Par exemple 10^{-3}.
- $\boldsymbol{\theta}$ Seuil, une valeur classique est $\frac{1}{\sqrt{n}}$.
- $\boldsymbol{\alpha}$ Pas d'apprentissage, $\alpha \in [0, 1]$. Une valeur faible ralentit l'apprentissage mais garantit une meilleure convergence.
- $\boldsymbol{\rho}$ Paramètre de vigilance. Permet de définir le nombre de catégories qu'il faut détecter. $\rho \in [0, 1]$, mais les meilleurs résultats sont pour $\rho \in [0.7, 0.9]$

b_{Ji} Les poids initiaux ascendants doivent être choisis de telle sorte qu'ils vérifient :

$$b_{Ji}(0) \leqslant \frac{1}{\epsilon \cdot \sqrt{n}}, \epsilon \in [0, 1[$$

afin d'éviter la détection d'un nouveau vainqueur au cours de la phase de résonance. Plus on augmente ϵ, plus le réseau aura tendance à former de nouvelles catégories.

t_{iJ} Les poids initiaux descendants doivent être proches de 0 afin de garantir qu'il n'y aura pas de remise à zéro lors de la présentation du premier patron. La valeur usuelle est 0.

$$f : f(x_i) = \begin{cases} x_i \; si \; x_i \leq \theta \\ 0 \; sinon \end{cases}$$

Algo *Art2*

Initialisation des paramètres : $\theta, \alpha, \rho, a, b, e, p$
ETAPE 1 : Répéter nbEpoch fois
  ETAPE 2 : $\forall \vec{s}$ faire les étapes 3 à 10
    ETAPE 3 : Mise à jour de la couche F1
      $u_i \leftarrow 0$; $w_i \leftarrow s_i$; $p_i \leftarrow 0$; $q_i \leftarrow 0$;
      $x_i \leftarrow \frac{s_i}{e+\|\vec{s}\|}$; $v_i \leftarrow f(x_i)$;
      $u_i \leftarrow \frac{v_i}{e+\|\vec{v}\|}$; $w_i \leftarrow s_i + a \cdot u_i$; $p_i \leftarrow u_i$;
      $x_i \leftarrow \frac{w_i}{e+\|\vec{w}\|}$; $q_i \leftarrow \frac{p_i}{e+\|\vec{p}\|}$;
      $v_i \leftarrow f(x_i) + b \cdot f(q_i)$;
    ETAPE 4 : Activation des y_j
      $y_j \leftarrow \sum_{i=1}^{n} b_{ij} \cdot p_i$;
    ETAPE 5 : Tant que raz Faire
      ETAPE 6 : $\forall J$ tq $y_J \geq y_j$, $\forall \; 1 \leq j \leq m$
        ETAPE 7 : Evaluer raz $u_i \leftarrow \frac{v_i}{e+\|\vec{v}\|}$;
        $p_i \leftarrow u_i + t_{Ji}$;
        $r_i \leftarrow \frac{u_i + p_i}{e+\|\vec{u}\|+\|\vec{p}\|}$;
        Si $\|\vec{r}\| < \rho - e$ Alors
        $j_J \leftarrow -1$; // inhibition de J
        $raz \leftarrow vrai$;
        Sinon
        $w_i \leftarrow s_i + a \cdot u_i$;
        $x_i \leftarrow \frac{w_i}{e+\|\vec{w}\|}$; $q_i \leftarrow \frac{p_i}{e+\|\vec{p}\|}$;
        $v_i \leftarrow f(x_i) + b \cdot f(q_i)$;
        $raz \leftarrow faux$;
        Fin si
    Fait // fin Etape 5
    ETAPE 8 : Répéter nbIt fois
      ETAPE 9 : MAJ des poids de la catégorie J
        $t_{Ji} \leftarrow \alpha \cdot u_i + t_{Ji}$;
        $b_{iJ} \leftarrow \alpha \cdot u_i + b_{iJ}$;
      ETAPE 10 : Mise à jour de F1
        $u_i \leftarrow \frac{v_i}{e+\|\vec{v}\|}$; $w_i \leftarrow s_i + a \cdot u_i$; $p_i \leftarrow u_i + t_{Ji}$;
        $x_i \leftarrow \frac{w_i}{e+\|\vec{w}\|}$; $q_i \leftarrow \frac{p_i}{e+\|\vec{p}\|}$;
        $v_i \leftarrow f(x_i) + b \cdot f(q_i)$;
    Fait // fin Etape 8
Fait // fin Etape 1

3 Apprentissage d'ART2

Nous avons créé différents réseaux en faisant varier les 2 paramètres principaux : la vigilance (ρ) et le pas d'apprentissage (α). La vigilance permet de reconnaître un mot nouveau quand il est présenté au réseau. Le pas d'apprentissage permet de déterminer si oui ou non il sera retenu par le réseau. Devant la durée nécessaire à l'entraînement d'un réseau, à cause de la taille et du nombre de vecteurs à traiter, suivant les paramètres utilisés, nous avons décidé de ne tester que quelques couples de valeurs :

$$(\rho, \alpha) \in \{0.3, 0.5, 0.7, 0.9\} \times \{0.4, 0.6, 0.8\}$$

soit $4*3 = 12$ couples par niveau. Nous avons donc obtenus $5 * 12 = 60$ réseaux (pour les 5 niveaux du CP au CM2) parmi lesquels nous avons sélectionné les deux meilleurs paramètres en terme de performance brute sur les mots fréquents, proches de celles décrites dans [6], ce qui donne $\rho \approx 0.7$ et $\alpha \approx 0.8$ comme nous le voyons en FIG. 1 puis dans la prochaine partie en FIG. 3.

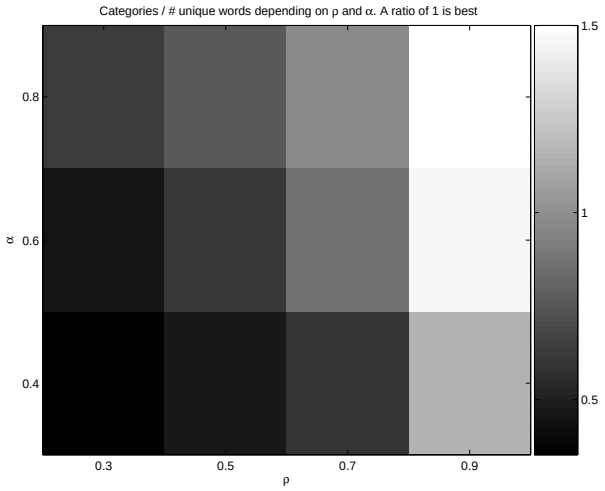

FIGURE 1. Catégories créées / # mots différents présentés. On cherche à s'approcher de 1. En abscisse ρ, en ordonnée α.

4 Résultats

4.1 Influence de ρ et α

En faisant varier ρ et α, nous pouvons observer leur impact sur l'apprentissage. Nous avons illustré avec les 30 premiers mots d'un manuel de CP ([11]), par ordre d'apparition, leur classification par 2 réseaux différents. Le Tab. 1 présente les mots par ordre de soumission au réseau. Ils sont très peu diversifiés (seulement 5 mots répétés) et sont donc intéressants pour cette expérience en offrant de fréquentes possibilités de rappels. ART2 devrait donc obtenir 5 categories soit une par mot.

La FIG. 2 montre cependant que le réseau ne sait pas reconnaître tous les mots : il identifie des mots nouveaux comme précédemment appris. Il en résulte un écrasement des catégories antérieures.

Mais la FIG. 3 montre un bon réseau. Une première confusion remplace *tralala* par *la* (mots 9 et 10) mais une nouvelle catégorie stable est crée pour l'occurence suivante

de *tralala* (mots 15, 26 et 30). Le réseau s'adapte donc et affine ses critères de classification (voir Tab. 2).

Les expériences montrent donc que c'est autour de $\rho \approx 0.7$ et $\alpha \approx 0.8$ que les résultats sont les plus pertinents. Les valeurs exactes seront à déterminer en fonction de nos exigences pour nous rapprocher du comportement humain observé [8].

1	gafi	**11**	la	**21**	est
2	tralala	**12**	la	**22**	gafi
3	tralala	**13**	la	**23**	gafi
4	tralala	**14**	est	**24**	est
5	tralala	**15**	tralala	**25**	moi
6	est	**16**	moi	**26**	tralala
7	moi	**17**	gafi	**27**	est
8	gafi	**18**	est	**28**	moi
9	tralala	**19**	gafi	**29**	gafi
10	la	**20**	moi	**30**	tralala

TABLE 1. Les 30 premiers mots de Gafi CP.

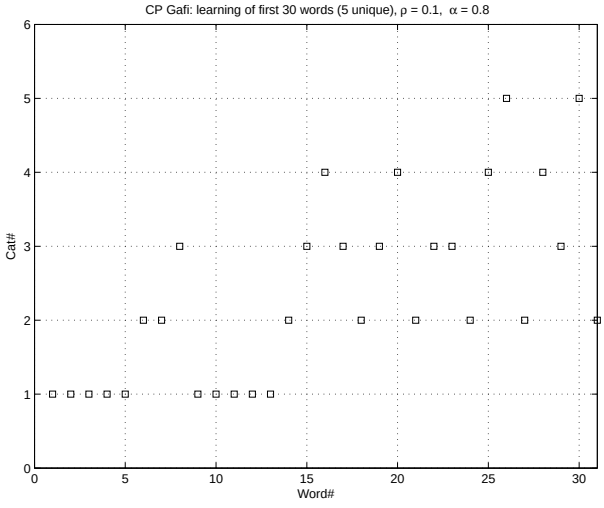

FIGURE 2. 30 premiers mots : $\rho = 0.1$ et $\alpha = 0.8$

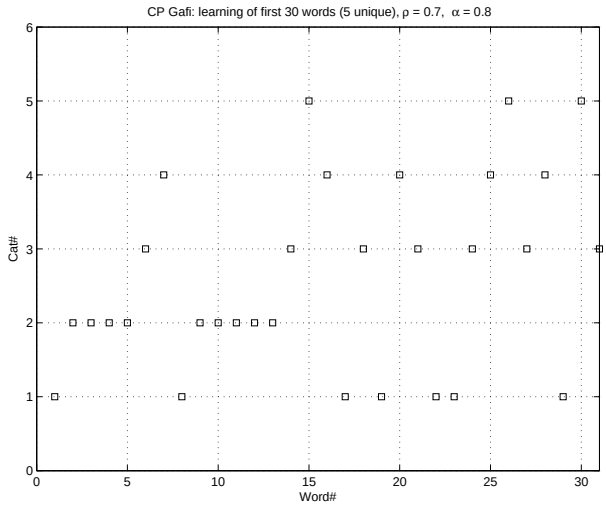

FIGURE 3. 30 premiers mots : $\rho = 0.7$ et $\alpha = 0.8$

Categorie	**1**	**2**	**3**	**4**	**5**
Mot	gafi	la	est	moi	tralala

TABLE 2. Contenu des catégories après l'apprentissage des 30 premiers mots pour : $\rho = 0.7$ et $\alpha = 0.8$

4.2 Mesures des perfomances du CP au CM2 sur séquences de mots originales ou aléatoires.

Nous avons étendu l'analyse sur tous les mots lus par un enfant (corpus réel) du CP au CM2 (5 niveaux). A chaque fin de niveau, nous mesurons pour une liste de mots ceux qui ont une catégorie propre. Cela conduit à un score de reconnaissance. Nous mesurons ce score pour une liste de 100 mots très fréquents dans les 5 corpus et de 100 mots très rares mais toutefois présents à au moins un exemplaire, dans chaque corpus (voir [6] pour détails).

Nous exécutons ART2 sur le corpus orginal et sur une variante avec sa séquence de présentation des exemples mélangée uniformément. En faisant varier le paramètre de vigilance, nous obtenons 2 modèles types : ART vigilant ($\rho = 0.7$) et ART moins vigilant ($\rho = 0.5$). Nous cherchons à savoir si l'uniformisation de la répartition des mots dans les corpus affecte plus un modèle que l'autre. Voici ce que nous obtenons :

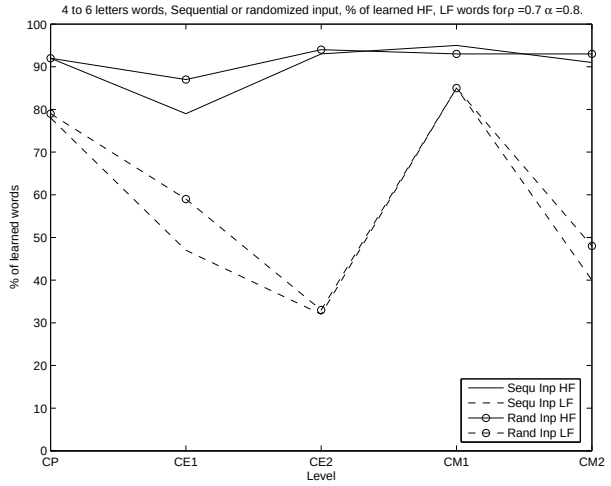

FIGURE 4. ART vigilant : Score de reconnaissance sur les listes de mots de 4 à 6 lettres très (HF) ou peu (LF) fréquents sur l'ensemble des 5 corpus, pour le réseau apprenant sur la séquence de mots dans l'ordre original (Sequential) ou mélangé uniformément (Random) ($\rho = 0.7, \alpha = 0.8$).

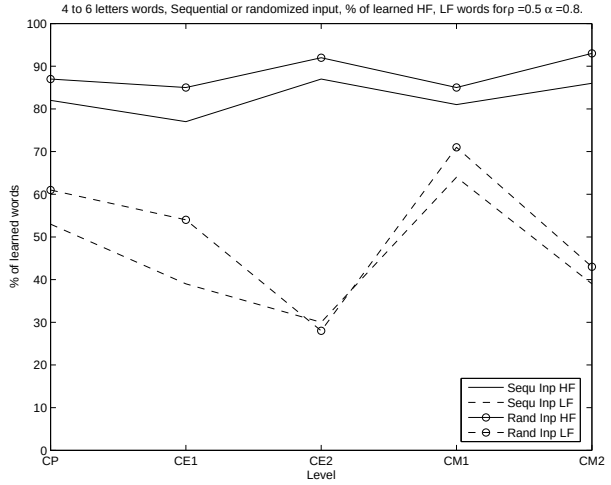

FIGURE 5. ART moins vigilant : Score de reconnaissance sur les listes de mots de 4 à 6 lettres très (HF) ou peu (LF) fréquent sur l'ensemble des 5 corpus, pour le réseau apprenant sur la séquence de mots dans l'ordre original (Sequential) ou mélangé uniformément (Random) ($\rho = 0.5, \alpha = 0.8$).

5 Discussion et conclusion

Nous avons pratiqué une analyse de variance (ANOVA) avec cinq facteurs indépendants :

- fréquence (2 niveaux : haute et basse)
- mode (2 niveaux : séquentiel et aléatoire)
- niveau (5 niveaux : 1 à 5)
- vigilance (4 niveaux : $0.3, 0.5, 0.7$ et 0.9)
- pas d'apprentissage (3 niveaux : $0.4, 0.6$ et 0.8).

Nous avons trouvé deux effets principaux. D'abord, nous avons mesuré un effet principal du facteur fréquence, $F(1, 208) = 217, p < 0.001$ indiquant que le modèle a appris plus de mots de haute fréquence (moyenne : 87.1) que de mots de basse fréquence (moyenne : 54.4). De plus, nous avons trouvé un effet principal du facteur vigilance, $F(1, 208) = 30, p < 0.001$, indiquant que le nombre de mots appris augmente avec la vigilance ($61.2, 69.4, 74.9$ et 77.5 mots en moyenne pour des niveaux de vigilance de $0.3, 0.5, 0.7$ et 0.9 respectivement).

Nous avons également trouvé des interactions significatives. La plus importante est une interaction du niveau par vigilance, $F(1, 208) = 12, p < 0.001$, illustrée en FIG.6. Comme nous pouvons le voir, le nombre de mots appris est anormalement bas au niveau 5 pour une vigilance de 0.9.

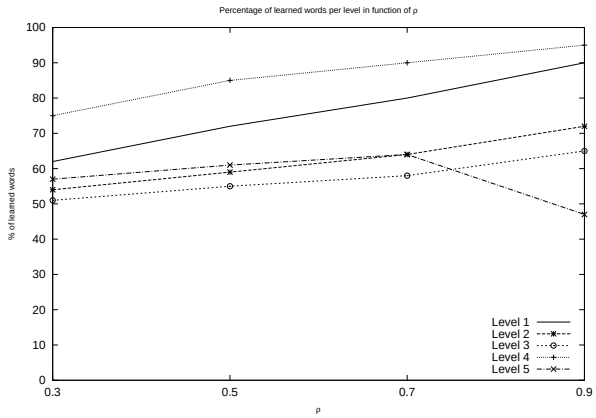

FIGURE 6. Pourcentage de mots appris par niveau en fonction de ρ.

Nous avons aussi trouvé une interaction entre la fréquence et la vigilance, $F(1, 208) = 4.3, p < 0.05$. Comme on le voit sur la FIG.7, le nombre de mots de basse fréquence appris est plus sensible aux variations de vigilance que les mots de haute fréquence.

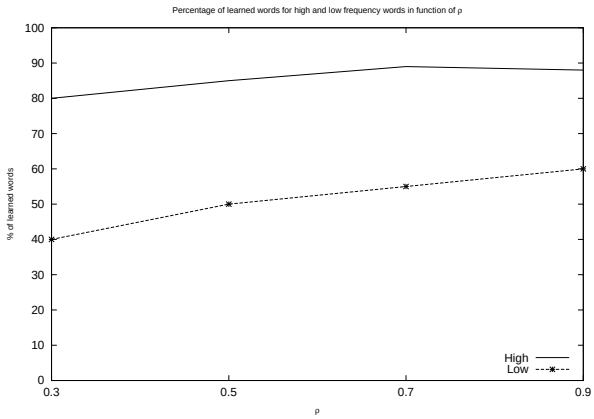

FIGURE 7. Pourcentage de mots appris pour les mots de haute et de basse fréquence en fonction de ρ.

Finalement, nous avons trouvé une interaction entre le niveau et l'apprentissage, $F(1, 208) = 4.7, p < 0.05$. Une inspection de la FIG.8 suggère que cette interaction est due au faible nombre de mots appris au niveau 3 lorsque la vigilance est de 0.8.

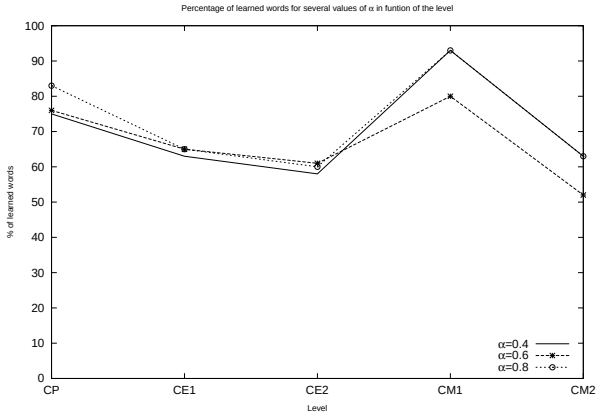

FIGURE 8. Pourcentage de mots appris pour différentes valeurs d'α en fonction du niveau.

Nous avons ensuite pratiqué une analyse de variance avec un facteur répété :

– niveau (5 niveaux : 1 à 5)

et deux facteurs indépendants :

– fréquence (2 niveaux : haute et basse)

– mode (2 niveaux : séquentiel et aléatoire).

Nous avons trouvé deux effets principaux. Premièrement, nous avons mesuré un effet principal du facteur fréquence, $F(1, 44) = 208, p < 0.001$ indiquant que le modèle a appris plus de mots de haute fréquence (moyenne : 87.1) que de mots de basse fréquence (moyenne : 54.4). Deuxièmement, nous avons trouvé un effet du facteur niveau $F(4, 176) = 71, p < 0.001$. Ce dernier est toutefois difficile à qualifier car le nombre de mots appris n'augmente pas de façon monotone en fonction du niveau. L'effet d'interaction niveau/fréquence est également significatif, quoique difficile à qualifier : $F(4, 176) = 44, p < 0.001$.

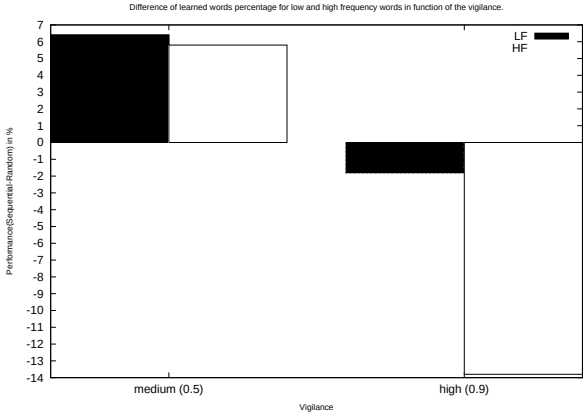

FIGURE 9. Différences des performances entre corpus séquentiel et aléatoire, pour les mots de basse fréquence LF (noir) ou haute fréquence HF (blanc), et pour une vigilance moyenne (gauche) ou forte (droite). Les valeurs positives correspondent à des performances supérieures sur les corpus séquentiels par rapport à celles sur les corpus aléatoires.

Afin d'évaluer l'influence de l'ordre de la présentation sur les performances en intéraction avec la vigilance du modèle, nous calculons pour deux types de modèles à vigilance moyenne (0.5) ou forte (0.9), les différences des performances moyennes sur les 5 niveaux, des mots BF ou HF. La FIG.9 montre que, quelque soit la fréquence des mots (5% ou 6%), l'ordre naturel est plus efficace pour un réseau moyennement vigilant. Pourtant, lorsque la vigilance augmente, la répartition aléatoire donne de meilleurs résultats pour les mots de haute fréquence (14%) alors que l'ordre n'affecte plus que très peu les mots de basse fréquence (2%).

En conclusion, les effets les plus importants sont (1) de fréquence : les mots de haute fréquence sont mieux appris que ceux de basse fréquence, et (2) de vigilance : plus de mots appris lorsque la vigilance est plus élevée. La fréquence est donc le facteur le plus important : le réseau apprend grâce à la répétition de la présentation des mots. Intuitivement, nous pouvons penser que les mots fréquents sont indépendants du contexte c'est à dire qu'ils sont distribués uniformément dans un corpus. Autrement dit, pour ces derniers, qu'un ordre séquentiel équivaut à un ordre aléatoire dans un corpus naturel. En effet, la FIG.9 montre une faible différence de taux de réussite pour le modèle moyennement vigilant (2^e colonne). Cependant pour un réseau très vigilant, les résultats montrent qu'une distribution aléatoire des mots induit une amélioration de l'apprentissage (4^e colonne). Donc ces résultats tendent à montrer qu'il est possible d'accroître significativement l'apprentissage des mots fréquents (donc usuels) en les répartissant plus uniformément dans les corpus. Les mots rares seraient

par contre dépendants du contexte : plusieurs mots rares peuvent par exemple apparaître à un seul endroit (imaginons un texte sur un thème particulier comme les "insectes" par exemple dans lequel une suite de mots rares seront utilisés et qui ne réapparaîtront plus dans la suite du corpus). La répartition aléatoire casse donc cette distribution contrairement à celle des mots fréquents. Ceci implique une interaction entre le type de présentation (en ordre ou mélangé) et la fréquence, avec un effet facilitateur plus marqué (FIG.9) sur les mots rares lors du mélange (car on permet aux occurrences de se distribuer uniformément dans le corpus). Malgré cela, la FIG.9 ne montre pas de nette amélioration de l'apprentissage en mélangeant uniformément les mots (colonnes 1 et 3). Notre choix d'hapax comme mots rares n'était peut-être pas judicieux et nous pourrions choisir des mots rares mais d'occurrences plus élevées. Cela pourrait accentuer les écarts entre l'ordre naturel et aléatoire. Finalement, nous pourrions aussi voir un effet plus significatif du facteur présentation en n'étudiant non plus l'apprentissage en fonction de la fréquence des mots dans les corpus mais leur voisinage orthographique ([6]).

6 Références

[1] Carpenter, G. A. and Grossberg, S., ART2 : Self-organization of stable category recognition codes for analog input patterns, *Applied optics*, 26, 1987, 4919-4930

[2] 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).

[3] 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).

[4] B. Lété, 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] Dufau, Lété, Touzet, Glotin, Ziegler, Grainger. Modélisation et simulation par carte auto-organisatrice de l'effet de frequence des mots chez l'apprenti lecteur. NEUROCOMP 2006

[7] Dufau, Grainger, Ziegler, Touzet, Lété & Glotin. Self-Organized Learning of Orthographic Representations. *Proceedings of the XVth Conference of the European Society for Cognitive Psychology (ESCOP 2007), Marseille, France.*

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

[9] 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.

[10] 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).

[11] M. Descouens, J-P. Rousseau. *Super Gafi CP - manuel élève* (France, Editions Nathan, 2003)