Ecrire pour le Web: ce que nous apprend la modélisation de la reconnaissance orthographique des mots

Stéphane Dufau, Claude Touzet, Jonathan Grainger

Version des auteurs non-éditée

Ecrire pour le Web:
ce que nous apprend la modélisation de la reconnaissance orthographique des mots

Stéphane Dufau [1], Claude Touzet [2] et Jonathan Grainger [1]

(1) Laboratoire de Psychologie Cognitive UMR6146
(2) Laboratoire de Neurobiologie Intégrative et Adaptative UMR6149

Résumé

Les travaux des ergonomes nous informent que des phrases courtes et percutantes sont mieux lues, et comprises, par les lecteurs internautes que des phrases longues et empesées. Cela ressemble à du bons sens et ne semble pas justifier de développements particuliers. Ce qui est peut être moins intuitif, mais tout aussi vérifié, c'est que ces phrases doivent utilisées le plus possible des mots fréquents, appris jeune et ayant peu de voisins orthographiques. Ces mots sont en effet reconnus plus rapidement que les autres. Le Webmaster curieux peut se demander d'où viennent ces effets de la fréquence d'occurrence, de l'âge d'acquisition, et du voisinage orthographique. Nous avons réalisé une modélisation neuronale du processus de reconnaissance orthographique des mots qui laisse supposer que le cerveau des lecteurs fonctionne comme une mémoire associative optimale. Les effets observés découlent principalement de la représentation neuronale des mots sous la forme d'un codage de type bigramme. Les paramètres influençant l'apprentissage de cette mémoire associative sont le nombre et l'ordre de présentation des mots. Ces paramètres sont principalement définis par les manuels scolaires et livres utilisés durant l'apprentissage de la lecture. Enfin, nous indiquons à partir de quelles ressources des outils informatiques peuvent être développés qui garantiraient pour chaque page Web un niveau d'accessibilité en terme de facilité de la reconnaissance orthographique.

1. Introduction

Les travaux des ergonomes nous informent que des phrases courtes et percutantes sont mieux lues, et comprises, par les lecteurs internautes que des phrases longues et empesées. Cela ressemble à du bons sens et ne semble pas justifier de développements particuliers. Ce qui est peut être moins intuitif, mais tout aussi vérifié, c'est que ces phrases (courtes et percutantes) doivent utilisées le plus possible des mots fréquents, appris tôt (dans la petite enfance par exemple) et ayant peu de voisins orthographiques. Un mot dont le nombre d'occurrences est élevé est reconnu plus rapidement chez le lecteur expert qu'un mot de plus faible occurrence (Monsell, 1991). Le voisinage orthographique d'un mot est défini comme un mot similaire à un autre à une lettre près (Coltheart, Davelaar, Jonasson & Besner, 1977). La densité de voisinage est le nombre de mots voisins. Un mot à forte densité est reconnu moins rapidement chez le lecteur expert qu'un mot à faible densité. Le Webmaster curieux peut se demander d'où viennent ces effets de l'âge d'acquisition, de la fréquence d'apparition et du voisinage orthographique. Nous avons réalisé une modélisation neuronale du processus de reconnaissance orthographique des mots qui laisse supposer que le cerveau des lecteurs fonctionne comme une mémoire associative optimale. Les effets observés découlent principalement de la représentation neuronale des mots sous la forme d'un codage de type bigramme. Le codage bigrammique des mots définit la notion de voisinage orthographique (quel mot est voisin de quel autre). Les paramètres influençant l'apprentissage de cette mémoire associative sont l'ordre de présentation des mots (i.e., l'âge d'acquisition) et le nombre de présentation d'un mot (i.e., la fréquence d'occurrence). Ces paramètres sont principalement définis par les manuels scolaires et livres utilisés durant l'apprentissage de la lecture (quels mots, dans quel ordre, avec quelle fréquence). Enfin, nous indiquons à partir de quelles ressources des outils informatiques peuvent être facilement développés qui fourniraient :

- le niveau de difficulté en reconnaissance orthographique pour chaque mot d'une page Web (calculé à partir de l'âge d'acquisition, la fréquence et le voisinage bigrammique),
- des mots synonymes de difficulté moindre pour chaque mot de difficulté élevée,
- un niveau d'accessibilité pour chaque page Web construit par exemple à partir de la somme des difficultés associées à chacun des mots de la page.

2. Modélisation neuronale de l'apprentissage de la reconnaissance orthographique des mots

2.1. La carte auto-organisatrice : une mémoire associative optimale

Nous avons choisi d'utiliser comme modèle neuronal de la reconnaissance orthographique des mots, la carte auto-organisatrice (CA) proposée par T. Kohonen (1982). Ce modèle est une tentative réussie de modélisation neuromimétique de l'organisation corticale. Par exemple, la CA explique de façon convaincante selon quel processus d'apprentissage auto-organisé se construisent les représentations somato-sensorielle et somato-motrice au niveau du cortex (Homonculus, fig. 1, Touzet, 1992). Le cortex est aussi le siège des fonctions cognitives impliquées dans la reconnaissance des mots.

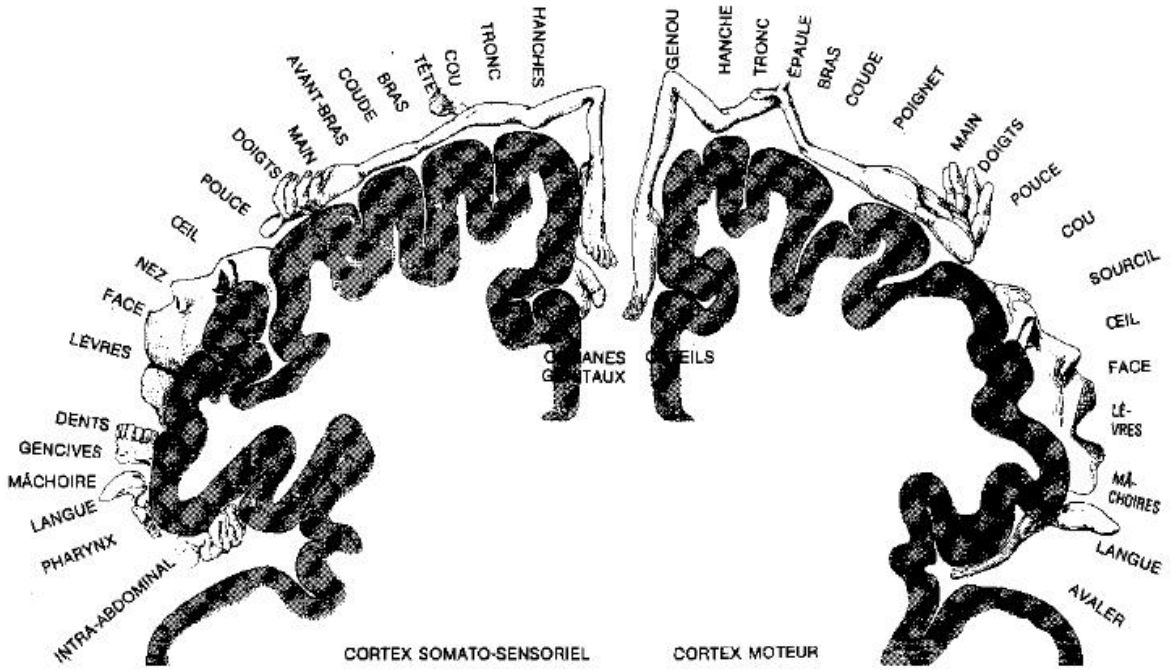

Figure 1. L'Homonculus (d'après Pour la Science, numéro spécial Le cerveau, 1982): la surface occupée au niveau cortical est fonction de la sensibilité sensorielle ou précision motrice de la partie du corps correspondante. Ainsi, la surface occupée par le pouce est supérieure à celle de la cuisse (arrangement spatial optimal). D'autre part, la topologie est conservée : les doigts sont l'un à coté de l'autre, etc.

La CA réalise une projection de l'espace d'entrée défini par les exemples d'apprentissage sur le sous-espace représenté par les neurones. Souvent, dans un but de visualisation, ce sous espace est de deux dimensions et justifie le nom de « carte » attribué à ce modèle. Cette carte peut être considérée comme optimale au sens ou la CA préserve la topologie et distribution de l'espace d'entrée lors de la projection. Les exemples proches dans l'espace d'entrée restent proches sur la carte, et les exemples plus fréquents sont plus représentés au niveau de la CA (fig. 2). Enfin la CA est une mémoire associative au sens où la position et la densité des exemples d'apprentissage sont mémorisées et peuvent être retrouvées pour peu que l'on fournisse une partie de l'information (quelques composantes du vecteur codant la position par exemple).

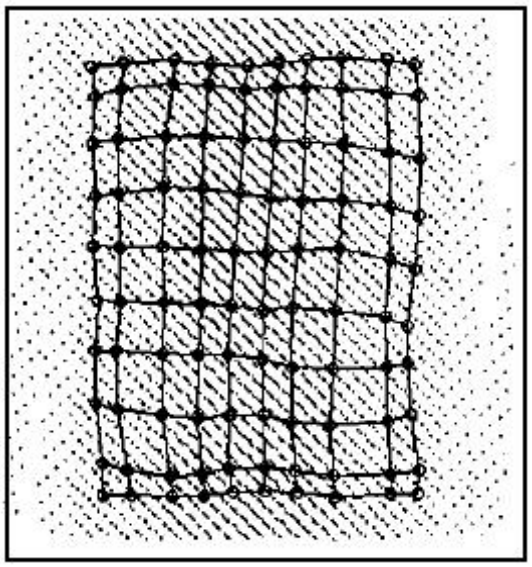

Figure 2. Occupation non uniforme d'un espace carré par un carte auto-organisatrice (d'après Touzet, 1992). On distingue les exemples d'apprentissage (les points, plus nombreux au niveau d'une bande verticale centrale), tandis que les 100 neurones de la carte sont indiqués par des cercles. Le voisinage entre les neurones est indiqué par des segments reliant les neurones voisins (4 voisins en moyenne par neurone). Les neurones se concentrent dans la zone de distribution la plus élevée. Il n'y a que 100 neurones pour représenter 3 000 exemples d'apprentissage : chaque neurone est donc représentatif en moyenne de 30 exemples (ou points).

2.2. Codage bigrammique des mots

Le codage bigrammique permet d'expliquer pourquoi une phrase comme "Sleon une édtue de l'Uvinertisé de Cmabrigde, l'odrre des ltteers dnas un mto n'a pas d'ipmrotncae" est aisément lisible et compréhensible (Grainger & Whitney, 2004). Cette phrase où les lettres sont mélangées nous indique que la position absolue des lettres est moins importante que leur position relative les unes aux autres. Grainger & van Heuven (2003) proposent de coder l'information de la position relative des lettres par un codage sous forme de bigrammes ouverts. Il s'agit en fait d'associer les lettres deux à deux même si ces lettres ne sont pas contiguës. Par exemple, CHUT est codé par -CH-CU-CT-HU-HT-UT- (fig. 3). Les bigrammes ouverts sont représentés par tous les bigrammes possibles. Il existe en français 41 lettres (lettres de A à Z, plus les lettres accentuées), c'est à dire 41 x 41 = 1681 combinaisons possibles. Un mot est donc une configuration de 1681 valeurs numériques (dans leur très grande majorité de valeur nulle).

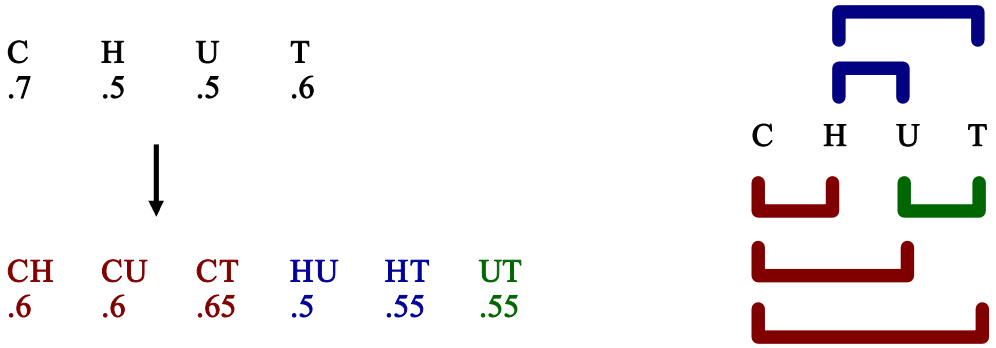

Figure 3. Codage du mot « chut » par des bigrammes ouverts (paires de 2 lettres). Une valeur numérique est associée à chacune des lettres qui traduit sa visibilité dans le mot (Peressotti & Grainger, 1999). Les bigrammes ouverts sont construits à partir des lettres et leur valeur numérique est la moyenne de la visibilité des lettres formant le bigramme.

2.3. Base d'apprentissage « naturelle »

Les mots utilisés pour l'apprentissage sont fournis à la CA dans leur ordre d'apparition au sein des manuels scolaires français (base lexicale Manulex; Lété, Sprenger-Charolles & Colé, 2004). La fréquence de répétition de ces mots dans la base d'apprentissage respecte celle rencontrée au sein des mêmes manuels. Nous utilisons dans nos simulations 855 mots différents pour l'apprentissage, leur nombre d'occurrences cumulées est de 8000.

3. Résultat de l'apprentissage par la carte auto-organisatrice

Comme nous le constatons à la lecture de la figure 4, la topologie de l'espace d'entrée est préservée : les mots orthographiquement proches se retrouvent proches sur la carte. De la même façon, le respect de la fréquence d'occurrence des mots est manifeste dans le fait qu'un même neurone correspond à un mot unique (haute fréquence) ou à plusieurs mots (tous de basse fréquence). Enfin, l'ordre d'apparition des mots est respectée au niveau de la carte : un mot présenté tardivement à peu de chances d'obtenir un neurone pour lui tout seul. Il est donc obligé de partager un même neurone avec d'autre(s) mot(s), ce qui gêne sa reconnaissance.

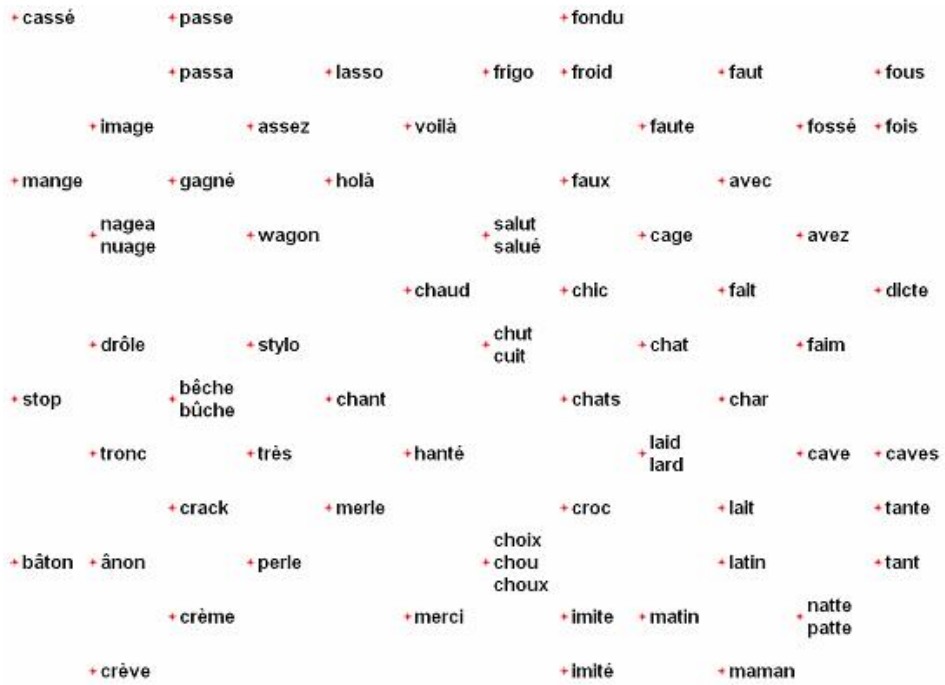

Figure 4. Une portion d'environ 13 x 13 neurones de la carte auto-organisatrice est représentée (taille totale 100 x 100). Certains neurones de la carte (points rouge) sont associés à un ou plusieurs mots, les autres neurones sont des voisins et codent donc pour des orthographes proches (mais ne correspondent pas à des mots existants). Les paramètres de l'algorithme d'apprentissage (Touzet, 1992) sont alpha = 0.5 et bêta = 0.05.

4. Comparaison avec des données expérimentales

Afin d'effectuer une comparaison de notre modèle avec les données expérimentales obtenues chez le lecteur expert (temps de réaction nécessaires à la reconnaissance d'un mot appris), nous définissons le « temps de réaction équivalent » à l'aide d'une fonction prenant en compte la distance avec les neurones voisins les plus proches dans les 4 directions (Nord, Sud, Est, Ouest), et le nombre de mots différents codés par le même neurone. Le temps de réaction équivalent est ainsi proportionnel à (N_mot + 0.25 x N_voisins) où N_mot est le nombre de mots codé par le neurone et N_voisins le nombre de neurones voisins occupés au moins par un mot. Nous utilisons pour cette comparaison (fig. 5) les mêmes bases d'exemples (mêmes manuels d'apprentissage et mêmes mots tests) pour les mesures effectuées chez le lecteur expert et avec notre modèle.
Dans l'expérience comportementale réalisée (tâche de démasquage progressif), les mots de bas voisinage orthographique (triangle) sont reconnus plus rapidement que ceux de haut voisinage (rond), et les mots de haute

fréquence (à droite) plus rapidement que ceux de basse fréquence (à gauche). La similitude des courbes présentées valide ainsi notre modèle.

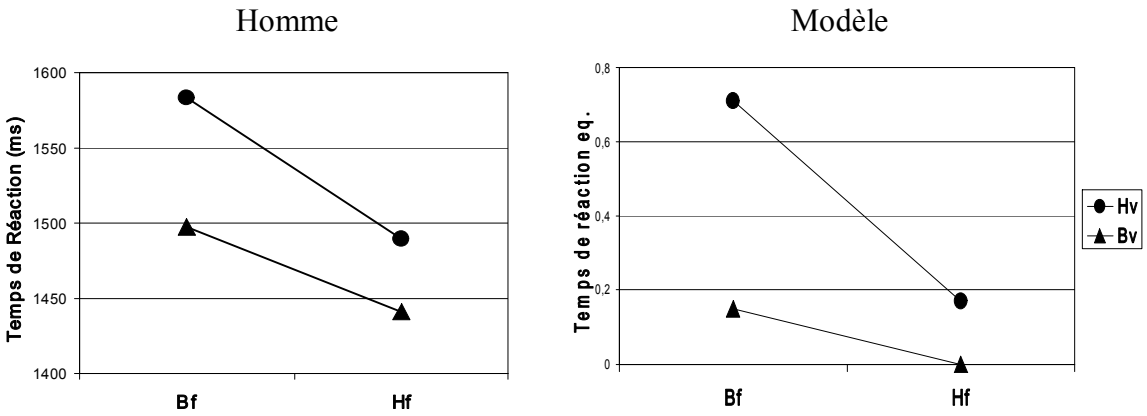

Figure 5. Comparaison des mesures de performances humaine et de performances du modèle sur 65 mots. L'apprentissage du réseau de neurones s'est fait sur 855 mots et 60 itérations d'apprentissage, paramètres alpha = 0.5 et bêta = 0.05.

5. Discussion

5.1. Voie orthographique vs. voie phonologique

La voie orthographique n'est qu'une des deux voies utilisées par le lecteur dans sa tâche de reconnaissance des mots. Notre modèle neuronal ne rend compte que de cette partie du traitement de l'information. Si Frost (1998) insiste sur le rôle de la phonologie dans la reconnaissance de mots écrits pour la plupart des expériences, la tâche de démasquage progressif privilégie cependant la voie orthographique. Il est néanmoins souhaitable de développer à l'avenir un modèle capable de tenir compte des deux voies orthographique et phonologique.

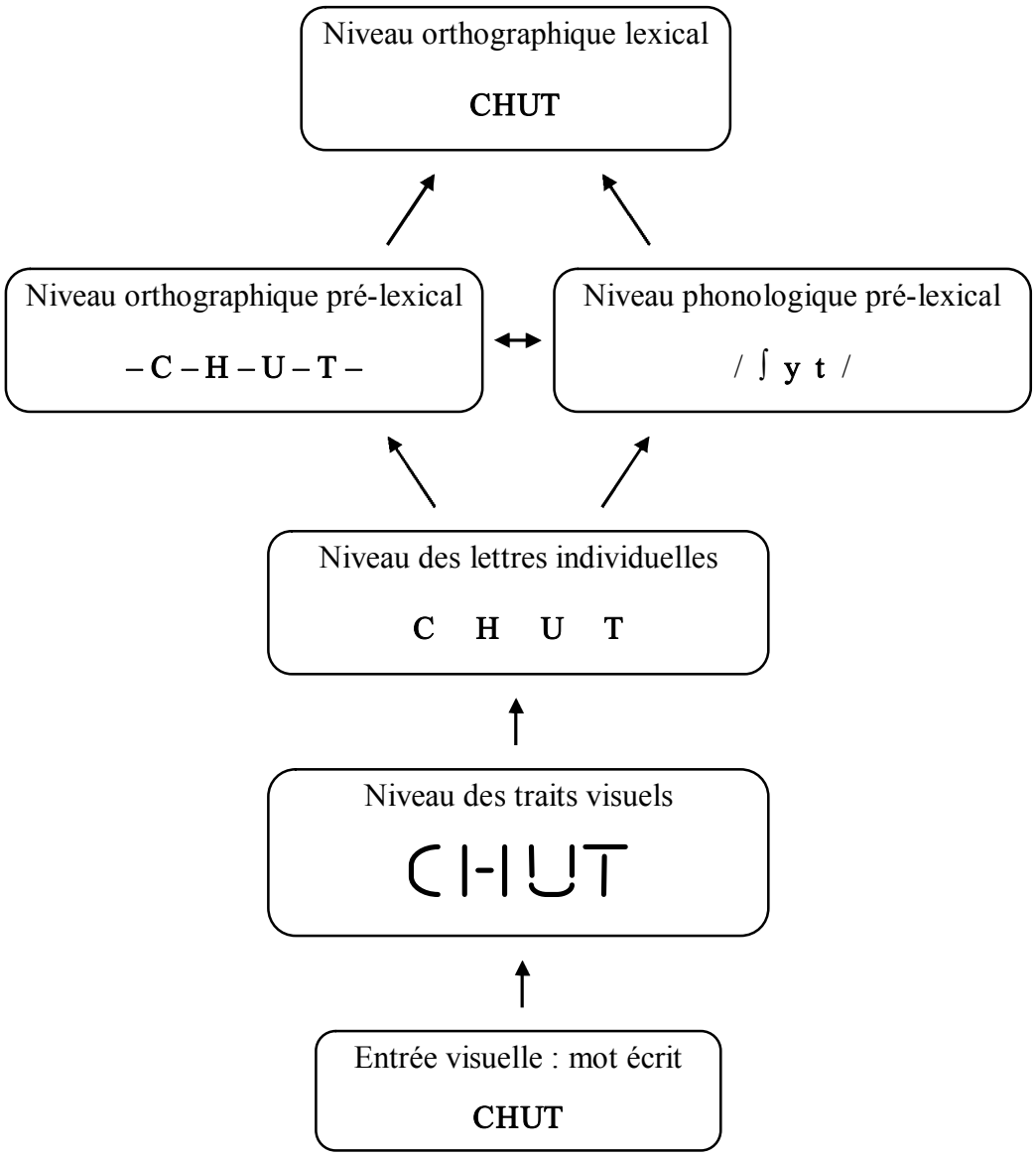

Figure 6. Modèle à deux voies de reconnaissance des mots écrits d'après Coltheart, Rastle, Perry, Langdon & Ziegler, 2001.

5.2. Age d'acquisition

Les enfants, surtout les plus jeunes sont particulièrement exposés aux effets répertoriés dans ce chapitre (Morrison & Ellis, 1995). La figure 7 montre que le différentiel observé chez les enfants tend à se réduire au cours de l'apprentissage. Cependant, certains effets persistent jusqu'à l'âge adulte : les mots de hautes fréquence sont reconnus plus rapidement.

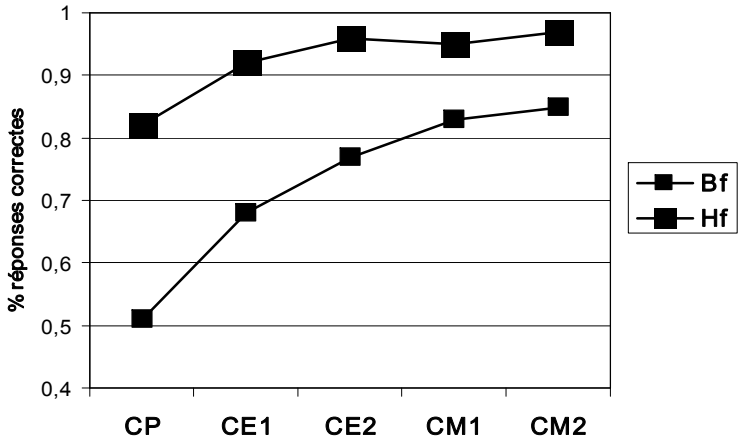

Figure 7. Taux d'identification correcte des mots au cours de l'apprentissage, selon leur fréquence (haute et basse).

6. Outils d'amélioration de l'accessibilité

Comme nous l'avons vu, la vitesse de reconnaissance orthographique n'est pas la même pour tous les mots, et ses effets sont maximum chez l'enfant. Certains dyslexiques (enfants et aussi adultes) rencontrent des difficultés encore plus importantes lors de la reconnaissance des mots. C'est pourquoi nous choisissons d'évoquer la notion d'accessibilité pour caractériser le niveau de difficulté que rencontre cette population lors de la reconnaissance orthographique des mots d'une page Web. Afin de garantir la meilleure accessibilité possible, il faut pouvoir mesurer celle-ci, et éventuellement pouvoir l'améliorer.

6.1 Mesure de l'accessibilité en terme de reconnaissance orthographique des mots

La base lexicale MANULEX (Lété et al., 2004) permet de réaliser des expériences sur un matériel linguistique contrôlé. Basée sur 54 manuels scolaires représentant un corpus de plus de 2 millions de mots, la base lexicale donne la fréquence d'occurrence des mots selon le niveau scolaire du CP au CM2. Cette base nous paraît être la plus représentative du développement des connaissances lexicales chez l'enfant. MANULEX en fournissant pour chaque mot sa fréquence, son nombre de voisins et son âge d'acquisition, permet de calculer le niveau de difficulté lors de sa reconnaissance orthographique. Il est ainsi possible de déterminer une note d'accessibilité de la page à partir de la somme des difficultés associées à chacun des mots de la page.

6.2. Modification de l'accessibilité en terme de reconnaissance orthographique des mots

Disposant du niveau d'accessibilité de chaque mot de la page, il est aisé alors de proposer, pour les plus difficiles d'entre eux, des synonymes dont on vérifiera en même temps le niveau d'accessibilité.

Les moyens informatiques d'aujourd'hui, à commencer par le Web et les systèmes de gestion de bases de données, en passant par les projets open-source et ceux développés par la « communauté », nous font croire que le développement de ces deux outils sera réalisé très prochainement.

Références

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), Attention and Performance: Vol. VI. Proceedings of the sixth international symposium on attention and performance, pp. 535-555. London: Academic Press.

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, 204-256.

Frost R. (1998). Toward a strong phonological theory of visual word recognition: true issues and false trails. *Psychoogical Bulletin,* 123(1), 71-99.

Grainger, J., & Van Heuven, W. (2003). *Modeling Letter Position Coding in Printed Word Perception*. In P. Bonin (Ed.), The Mental lexicon, pp. 1-24. New York : Nova Science Publishers.

Grainger, J,. & Whitney, C. (2004). Does the huamn mind raed wrods as a wlohe? *Trends in Cognitive Sciences*, 8, 58-59.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary-school readers . *Behavior Research Methods, Instruments, & Computers, 36*, 156-166.

Morrison C., & Ellis A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 116-133.

Monsell S. (1991). *The nature and locus of word frequency effects in reading*. In D. Besner (Ed) & G. Hunphreys (Ed), Basic processes in reading: Visual word recognition, pp. 148-197. Hillsdale, NJ (Lawrence Erlbaum Associates).

Peressotti, F., & Grainger, J. (1999). The role of letter identity and letter position in orthographic priming. *Perception & Psychophysics*, 61, 691-706.

Remerciements

Les auteurs remercient :

Hervé Glotin (Laboratoire des Sciences de l'Information et des Systèmes, Toulon),
Bernard Lété (Laboratoire d'Etude des Mécanismes Cognitifs , Lyon),
Johannes Ziegler (Laboratoire de Psychologie Cognitive, Marseille),

partenaires du programme interdisciplinaire du CNRS – Actions concertées incitatives (Traitement des Connaissances, Apprentissage et NTIC). *Modélisation computationnelle de l'apprentissage des mots écrits.*