Quand la syntaxe guide la compréhension en lecture

Premières approches issues du modèle structural de lecture.

Denis Foucambert

Département de linguistique et de didactique des langues

Université du Québec à Montréal

foucambert.denis@uqam.ca

INTRODUCTION

Le rôle de la syntaxe passe le plus souvent au deuxième plan des descriptions de l'acte lexique, aussi bien en ce qui concerne le lecteur expert (Posner, Abdullaev, McCandliss, Sereno, & Everatt, 1999), que pour l'apprentissage de la lecture. L'idée d'une analyse syntaxique « on line », dépendante d'une identification des mots par un accès au lexique mental, et séquentielle, dans la mesure où elle s'effectue pas à pas au fur et à mesure que les mots sont identifiés (Lecocq, Casalis, Leuwers, & Watteau, 1996) est le plus souvent décrite dans la littérature.

Néanmoins, depuis 15 ans, le modèle structural de lecture (Greenberg, Healy, Koriat, & Kreiner, 2004; Greenberg & Koriat, 1991; Koriat & Greenberg, 1996) se met progressivement en place ; il postule que le lecteur tente d'établir un cadre structural de la phrase (ou du groupe de mots) à l'aide du repérage des unités organisant la syntaxe pour ensuite y intégrer les différentes informations sémantiques. La lecture experte ne relèverait donc pas d'abord d'une activité séquentielle qui implique l'identification des mots les uns après les autres afin de leur attribuer un rôle (syntaxique ou non) mais serait "pilotée" – en partie à l'aide des informations perçues à la périphérie de la rétine (Saint-Aubin & Klein, 2001) – par le repérage d'unités syntaxiques qui organisent le contenu sémantique de la phrase parcourue.

Le paradigme expérimental utilisé majoritairement pour mettre au point ce modèle repose sur l'observation des oublis dans une tâche de détection de lettres sur un texte : la différence d'oubli de lettres entre celles présentes dans les mots organisant la syntaxe et celles présentes dans les mots lexicaux est expliquée par le rejet à l'arrière-plan des mots « structurants » tandis que l'intégration du sens se poursuit à l'aide des mots « lexicaux », dans le cadre structural pré-construit. Cette tâche expérimentale, largement utilisé en psychologie, a permis le développement de théories diverses expliquant de manière contradictoire l'oubli de lettres : le modèle structural de lecture s'est développé en opposition avec le modèle d'unités concurrentes, qui explique l'oubli par la reconnaissance rapide

(globale) d'un mot fréquent, ce qui stoppe les niveaux d'analyse parallèles comme la reconnaissance des lettres, et provoque donc un taux plus fort de l'oubli de lettres dans les mots les plus fréquents (Healy, 1994).

Si les deux modèles sont aujourd'hui unifiés, avec leur rôle et leur temporalité respectif, au sein d'un schéma explicatif unique – le modèle Guidance-Organisation (GO) développé par les principaux auteurs des deux modèles précédemment concurrents (Greenberg et al., 2004), il n'en reste pas moins que la partie structurale du modèle GO reste non seulement d'actualité, mais qu'elle est toujours considérée comme première dans la temporalité des processus, avec une activation rapide du squelette syntaxique de la phrase ou du groupe de mots.

Cependant, la totalité des expérimentations n'a jusqu'alors porté que sur des publics ayant un niveau de lecture non contrôlé, généralement étudiants en université, comme si cette population était homogène dans ses performances en lecture. Notre question de recherche vise à mesurer des différences d'habileté syntaxique en fonction des différences interindividuelles de lecture, en faisant l'hypothèse générale que, à des différences interindividuelles de vitesse et de compréhension en lecture correspondent des différences significatives dans l'oubli de lettres entre les mots qui organisent la syntaxe de la phrase et les mots plus chargés sémantiquement.

MÉTHODE

Population

Soixante-quatre adultes, tous enseignants, donc au moins titulaires du baccalauréat français, vont passer l'ensemble des épreuves que nous allons décrire. L'âge moyen est de 40,7 ans, avec un maximum de 58 ans et un benjamin de 26 ans. L'âge médian est de 41 ans. La distribution en âge de la population est normale (Test de Kolmogorov-Smirnov $d = 0,071$, $p>0,20$). On dénombre 16 hommes

pour 48 femmes. Cette répartition 25% / 75% ne saurait étonner dans le monde enseignant où la surreprésentation féminine est la règle.

Épreuves

A. Les épreuves de lecture

Pour évaluer la compréhension et la vitesse, nous nous sommes servis de deux épreuves différentes, déjà utilisées dans de précédents travaux (Foucambert, 2000, 2003). Elles nous semblent révélatrices de ce qui attend un lecteur dans sa pratique quotidienne : parcourir des textes relativement simples pour en extraire de l'explicite et conduire une lecture savante sur un texte.

La première épreuve tente d'évaluer la performance dans une forme de lecture très courante, sans doute à l'œuvre dans plus de 60 % des situations ordinaires, celle où il s'agit de prendre connaissance simplement de l'explicite d'un texte, ce qui correspond à ce que l'ex Direction des Études et Prospective du Ministère de l'Éducation Nationale décrivait comme une "compétence approfondie" (Vugalic, 1996). L'épreuve se déroule sur ordinateur dont la résolution d'écran est contrôlée (800*600). Chacun des neuf textes s'affiche ; le sujet indique, en pressant une touche du clavier, qu'il en a terminé la lecture et répond alors à une question. Les textes sont diversifiés entre presse, documentaire et fiction, sont d'une taille similaire d'environ 20 lignes et de même niveau de difficulté aussi bien pour le lexique employé que pour la complexité des phrases. Les questions portent sur des points explicitement présents dans le texte et sont systématiquement introduites par la formule « Le texte parle :» suivie de trois propositions parmi lesquelles une seule est correcte.

Une seconde épreuve fait travailler sur l'implicite du texte, ce que l'ex DEP dénommait "compétence remarquable". Il s'agit de franchir ce que dit le texte pour atteindre l'intention de l'auteur et apprécier les moyens qu'il emploie. Un texte de fiction, long de 1526 mots, est présenté sur un écran d'ordinateur ; ce texte permet de nombreuses interprétations, en partie par l'usage que fait l'auteur de différents épilogues. Le sujet peut parcourir à sa guise les 9 pages écran pendant le temps qu'il estime

nécessaire. Ensuite, il répond à 12 questions par un système de QCM, le texte n'étant alors plus consultable. Un barème a été établi par un groupe de juges formés d'enseignants et de bibliothécaires pour décrire des degrés d'interprétation et ne pas s'enfermer dans le tout ou rien. Ainsi, toutes les réponses proposées sont possibles mais certaines témoignent d'un niveau plus approfondi de compréhension qui est calculée en s'efforçant de correspondre aux conceptions de la lecture experte d'un texte littéraire.

B. Les tâches de détection de lettres

Quatre textes supports sont présentés sur quatre jours aux sujets (un par jour). Les textes comportent entre 421 et 590 mots et les lettres à détecter sont successivement le *u*, le *d*, le *r* et le *t*. Nous avons choisi de neutraliser la différence d'oubli en fonction de la lettre, souvent observée, en présentant plusieurs lettres de visibilité différente (Nazir, Jacobs, & O'Regan, 1998). Les textes sont présentés de manière identique, dans un double interlignage, dans la fonte Times New Roman avec une taille de 12 points. Sur les 2044 mots de l'ensemble des textes, on en compte 617 contenant les différentes lettres à barrer.

Mesures

A. Les épreuves de lecture

Pour chacune des épreuves de lecture, nous gardons deux variables primaires, à savoir la vitesse de lecture et la compréhension (voir tableau 1). La première est calculée en mots par heure, tandis que la deuxième est une note sur 100.

INSÉRER TABLEAU 1

Deux questions se posent pour extraire de ces quatre variables primaires des informations pertinentes. La première est de savoir comment rendre complètement indépendantes les vitesses et les

compréhensions de chacun des individus. Cette séparation s'impose car nous chercherons, dans la suite de ce travail, à mettre en rapport la vitesse et la compréhension avec l'oubli de lettres. La deuxième est de s'assurer que les phénomènes illustrés par les valeurs numériques sont suffisamment homogènes. Pour répondre à ces deux impératifs, il nous a semblé qu'une Analyse en Composante Principale était une méthode appropriée. Plutôt que de simplement additionner les résultats obtenus par les sujets, nous avons construit différents indices à partir des axes d'une Analyse en Composantes Principales. Ce type d'analyse factorielle est en effet particulièrement utile quand on veut synthétiser plusieurs variables à partir des liaisons qu'elles entretiennent. Ainsi, l'analyse en composantes principales en éliminant successivement les corrélations multiples entre les résultats aux différentes épreuves permet de décomposer chaque épreuve en fonction des liens multiples que son résultat entretient avec les autres. Cette analyse aboutit à remplacer ici les résultats des 2 épreuves par 4 composantes plus ou moins simultanément à l'œuvre dans chaque résultat.

Les quatre variables primaires issues des deux épreuves sont les variables actives qui permettent de construire les 4 facteurs de l'ACP, décrits par le tableau 2.

INSÉRER TABLEAU 2

Les trois premiers axes de l'analyse rendent compte, à eux seuls, de plus de 95% de la variance totale ; la figure 1 les présente.

INSÉRER FIGURE 1

On remarque que le premier axe est construit par les deux vitesses de lecture alors que le deuxième l'est par les deux scores de compréhension. Le troisième axe, quant à lui, oppose les compréhensions issues des deux épreuves. En conséquence, l'espace proposé par les trois premiers axes représente bien la compétence générale en lecture, en tenant compte aussi bien des phénomènes

qui unissent les vitesses de lecture (axe 1) que de ceux qui rapprochent les processus à l'œuvre dans les deux compréhensions (axe 2). Le troisième axe différencie les processus de compréhension entre ceux assez factuels d'un texte simple et ceux plus profonds d'un texte à plusieurs niveaux de lecture (axe 3) ; entre la compréhension en lecture d'information (documentaire et presse) et en lecture littéraire, entre saisie de l'explicite et traitement de l'implicite. La compréhension est donc évaluée à l'aide de deux composantes de poids assez proches (26 et 23 % de variance exprimée) : l'axe 2 exprime ce qu'elles ont en commun et permet de parler d'un niveau général de compréhension tandis que l'axe 3 exprime ce qui les différencie selon la nature des textes qu'il s'agit de comprendre. Cette dissociation rejoint la complexité décrite par Fayol au sujet des processus de compréhension : « L'immense complexité des processus, où tout se modifie simultanément et en interaction : les signifiants, les signifiés, les procédures, les capacités de contrôle, etc… » (Fayol, 1992).

Dans leur ensemble, les deux phénomènes vitesse et compréhension ont un poids similaire, puisque l'axe 1 explique 44 % de la variance alors que les deux axes représentant la compréhension en expliquent un peu plus de 50%.

L'axe 1 représentant massivement la vitesse de lecture, la coordonnée des individus sur cet axe sera réutilisée pour définir leur vitesse de lecture dans les analyses ultérieures. De la même façon et pour des raisons similaires, les coordonnées des individus sur l'axe 2 et sur l'axe 3 définiront respectivement leur niveau et leur style de compréhension. Dorénavant, nous appellerons la compréhension construite par l'axe 2 de l'ACP la compréhension^{U} (pour Union) et celle construite par l'axe 3 la compréhension ^{D} (pour Différence).

B) La détection de lettres

La variable calculée est le rapport entre les omissions dans les mots à rôle syntaxique et celles dans les mots à rôle sémantique. Ce rapport est calculé sur une observation d'environ 40000 mots[1]. Le regroupement des déterminants, prépositions et complémenteurs constituent les mots à rôle syntaxique ; les mots à rôle sémantique sont les noms, les adjectifs, les adverbes et les verbes. Cette

variable (m = 3,5 ; σ = 2,09 ; min. = 0,75 ; max. = 10,34) présente une distribution normale (Kolmogorov-Smirnov : d = 0,12 ; p>0,20). Dans la suite, cette variable sera nommée V_{oubli}.

Résultats

Ce qui nous intéresse, c'est de savoir si la vitesse de lecture ou une des modalités de la compréhension peut s'expliquer par l'effet d'oubli de lettres. Pour répondre à cette question, nous procédons à une analyse en régression multiple. Nous introduisons dans le modèle les trois variables issues de l'Analyse en Composantes Principales comme variables dépendantes et les variables sexe, pourcentage général d'oublis et différentiel d'oubli seront nos trois variables indépendantes. Le modèle regroupe donc trois analyses de régression multiple. Le tableau 3 présente les résultats généraux du modèle.

INSÉRER TABLEAU 3

On remarque que le modèle explique deux variables dépendantes : la vitesse et la compréhension^{U}. Rappelons que la compréhension^{U} illustre les phénomènes de compréhension quel que soit le type de lecture, à l'inverse de la compréhension^{D} qui oppose ce qui se passe entre lecture documentaire et lecture littéraire. La compréhension^{U} semble illustrer des phénomènes plus généraux, à l'œuvre dans toute compréhension.

Le tableau 5 présente les coefficients de régression, les coefficients standardisés, et leur significativité, pour les variables continues ou les modalités.

INSÉRER TABLEAU 4

On observe que :

- la variable sexe intervient de manière significative dans la vitesse de lecture avec une vitesse de lecture plus importante pour les hommes[2].
- le pourcentage général d'oubli explique significativement la compréhension^{U}, avec un oubli général plus important pour les meilleurs compreneurs.
- le différentiel d'oublis explique, lui aussi, la compréhension^{U} avec un différentiel plus élevé pour les meilleurs compreneurs.
- la compréhension^{D} n'est expliquée par aucune des variables introduites dans le modèle.

DISCUSSION

Cette étude fait apparaître une relation significative entre oubli privilégié de lettres dans les mots organisant la syntaxe et compréhension de textes. Conformément à l'hypothèse structurale, le différentiel entre oublis de lettres dans les mots à rôle syntaxique et oublis dans les mots à rôle sémantique s'explique par le décalage entre le repérage de la syntaxe et le traitement des mots à rôle sémantique qui prennent place dans des structures anticipées. Le modèle statistique expliquant les performances en lecture montre qu'une meilleure compréhension^{U} est liée à un plus fort différentiel d'oubli. Il ne faut pas s'étonner que le pourcentage de variance expliquée pour la compréhension^{U} ne soit que de 14% : d'autres éléments contribuent bien évidemment à la compréhension, comme la familiarité avec le sujet traité, l'empan de mémoire de travail (Daneman & Carpenter, 1980),… Cependant, les résultats montrent sans conteste que le bon compreneur est celui qui, à vitesse égale, anticipe avec plus de sûreté la construction de la phrase, ce dont témoigne le rapport entre les taux d'oubli de lettres dans les mots à rôle syntaxique et celui dans les mots à rôle non syntaxique.

En outre, il est important de souligner la corrélation négative entre le différentiel d'oublis et le taux général d'oubli de lettres : de manière générale, les lecteurs qui établissent efficacement ou tentent

d'établir des cadres structuraux, ce qui se repère à l'aune de leur différentiel, sont en même temps ceux qui sont le plus attentifs au matériau graphique proposé.

La corrélation positive entre la compréhension en lecture et la capacité à avoir construit l'organisation syntaxique avant le traitement des mots semble contredire certaines descriptions classiques de l'acte lexique et confirment certains développements du modèle structural qui considèrent que la formation de la structure syntaxique de la phrase (ou d'un groupe de mots) est édifiée a priori à l'aide d'une prise d'information parafovéale. A partir de nos résultats, nous pouvons affirmer que le meilleur compreneur est celui qui, à vitesse égale, anticipe avec le plus de créativité les constructions syntaxiques : le lien entre la compréhension^{U} et un différentiel d'oubli important suggère que l'habileté dans la construction des cadres syntaxiques (qui vont être remplis ensuite par un contenu sémantique) est prédictif d'une bonne compréhension. A l'inverse, les moins bons compreneurs, ayant une habileté syntaxique moindre, doivent mener de front, dans leurs rencontres avec les différents mots de la phrase, la construction de la syntaxe et celle du sens.

Ces résultats ouvrent la voie à des recherches plus didactiques où il sera nécessaire de vérifier si l'entraînement de cette habileté provoque une amélioration de la compréhension en lecture.

Bibliographie

Daneman, M., & Carpenter, P. A. (1980). Individual difference in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Fayol, M. (1992). Comprendre ce qu'on lit: de l'automatisme au contrôle. In M. Fayol, J. E. Gombert, P. Lecocq, L. Sprenger-Charolles & D. Zagar (Eds.), *Psychologie cognitive de la lecture.* (pp. 73-106). Paris: PUF.

Foucambert, D. (2000). Les effets d'une année d'entraînement à la lecture avec un logiciel éducatif : résultats en classe de sixième de collège. *Revue Française de Pédagogie, 133*, 63-73.

Foucambert, D. (2003). *Syntaxe, vision parafovéale et processus de lecture. Contribution du modèle structural à la pédagogie.* Unpublished Thèse de doctorat en sciences de l'Éducation., Université Grenoble 2.

Greenberg, S. N., Healy, A. F., Koriat, A., & Kreiner, H. (2004). The GO model: A reconsideration of the role of structural units in guiding and organizing text on line. *Psychonomic Bulletin and Review, 11*(3), 428-433.

Greenberg, S. N., & Koriat, A. (1991). The missing-letter effect for common function word depends on their linguistic function in the phrase. *Journal of Experimental Psychology : Learning, Memory and Cognition, 17*, 1051-1061.

Healy, A. F. (1994). Letter detection : a window to unitization and other cognitive processes in reading texts. *Psychonomic Bulletin and Review, 1*, 333-344.

Koriat, A., & Greenberg, S. N. (1996). The Enhancement Effect in Letter Detection : Further Evidence for the Structural Model of Reading. *Journal of Experimental Psychology : Learning, Memory and Cognition, 22*, 1184-1195.

Lecocq, P., Casalis, S., Leuwers, C., & Watteau, N. (1996). *Apprentissage de la lecture et compréhension d'énoncés*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Nazir, T. A., Jacobs, A. M., & O'Regan, J. K. (1998). Letter legibility and visual word recognition. *Memory & Cognition, 26*(4), 810-821.

Posner, M. I., Abdullaev, Y. G., McCandliss, B. D., Sereno, S. C., & Everatt, J. (1999). Anatomy, circuitry and plasticity of word reading. In *Reading and Dyslexia: Visual and Attentional Processes* (pp. 137-163). London: Routledge.

Saint-Aubin, J., & Klein, R. M. (2001). Influence of parafoveal processing on the missing-letter effect. *Journal of Experimental Psychology : Learning, Memory and Cognition, 27*(2), 318-334.

Vugalic, S. (1996). *Les compétences en lecture, en calcul et en géométrie des élèves … l'entrée au CE2 et en sixième* (No. Note d'information, 96.22.): Ministre de l'Éducation nationale, de l'enseignement supérieur et de la recherche.

TABLEAUX ET FIGURES

Tableau 1 – Les résultats primaires des deux épreuves de lecture

	Vitesse (m/h)		Compréhension	
	Textes courts	Texte long	Textes courts	Texte long
Moyenne	18202	20593	78,9	76,7
σ	5600	5701	21,7	9,8
Minimum	11659	10258	0	40
Maximum	40128	35345	100	91

Tableau 2 – Résultats de l'ACP : Variance expliquée et valeurs propres de chacun des axes.

Axe	**Valeur propre**	**Pourcentage de la variance expliquée**	**Valeur propre cumulée**	**Pourcentage cumulé**
1	1,790168	44,75420	1,790168	44,7542
2	1,071344	26,78359	2,861512	71,5378
3	0,945825	23,64562	3,807337	95,1834
4	0,192663	4,81659	4,000000	100,0000

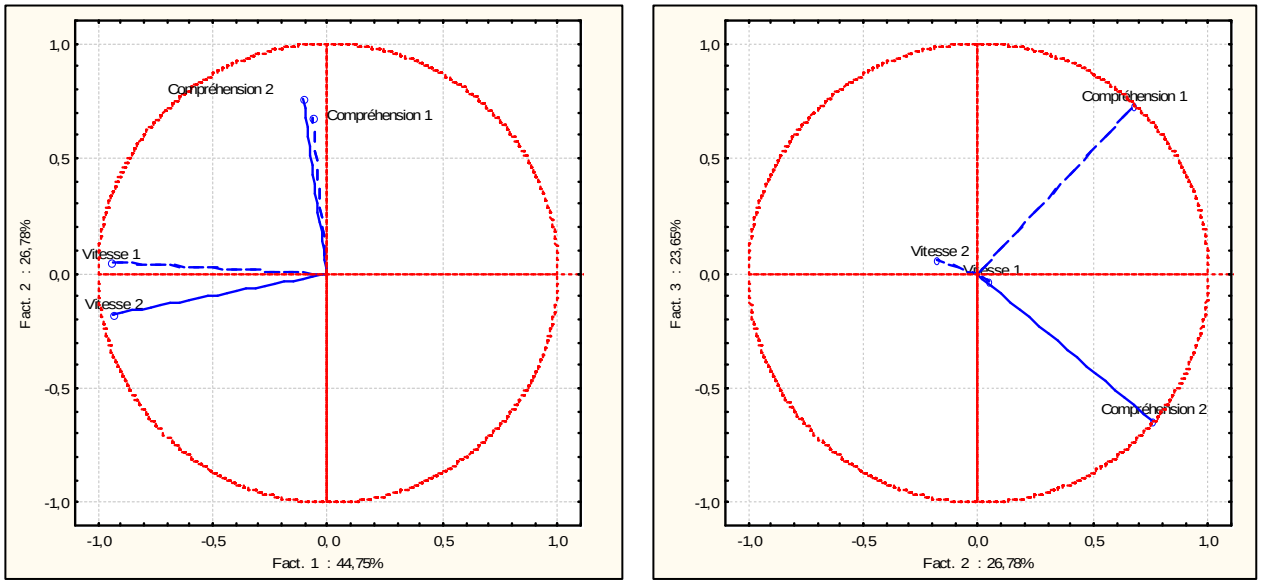
Tableau 3 – Résultats généraux de la régression cherchant à expliquer les résultats en lecture par l'oubli de lettres.

Variables dépendantes	**Corrélation**	**Pourcentage de variance expliquée**	**p**
Vitesse	**0,42**	**18%**	**<0,01**
Compréhension ^{U}	**0,38**	**14%**	**<0,03**
Compréhension ^{D}	0,09	0,7%	<0,93

Tableau 4 – Paramètres, coefficients bêta et significativité des variables explicatives du modèle expliquant les performances en lecture.

			Ord. Origine	**Variables explicatives**		
				Sexe	**Pourcentage général d'oubli**	**Différentiel d'oubli**
Variables à expliquer	**Vitesse de lecture**	Paramètre	-0,41	**-0,61**	0,009	-0,03
		Coeff. bêta		**-0,4**	0,08	-0,05
		Significativité	0,58	**<0,001**	0,60	0,72
	Compré-hension ^{U}	Paramètre	-1,7	0,1	**0,03**	**0,2**
		Coeff. bêta		0,11	**0,43**	**0,45**
		Significativité	<0,01	<0,39	**<0,01**	**<0,01**
	Compré-hension ^{D}	Paramètre	0,18	0,03	-0,0008	-0,04
		Coeff. bêta		-0,03	-0,01	0,09
		Significativité	<0,8	<0,85	<0,96	<0,61

Figure 1 — Facteur 1, 2 et 3 de l'ACP sur les épreuves de lecture (Vitesse 1 et Compréhension 1 pour l'épreuve 1 ; Vitesse 2 et Compréhension 2 pour l'épreuve 2)



NOTES

[1] On compte 617 mots contenant des lettres à détecter et 64 sujets : la multiplication donne mots 39488 occurrences.
[2] Le paramètre est certes négatif, mais la vitesse de lecture est issue de l'axe 1 de l'ACP qui situait les vitesses les plus élevées à son côté négatif.