Rôle de l'information visuelle dans l'accès au lexique

Mathilde Fort[a], Justine Chipot[a], Sonia Kandel[a&b], Christophe Savariaux[c] et Elsa Spinelli[a&b]

[a]Université Pierre Mendès France, Laboratoire de Psychologie et NeuroCognition (CNRS UMR 5105)
BP 47 – 38040 Grenoble Cedex 9, France Tel.: +33 4.76.82.56.30 ; Fax : +33 4.76.82.78.34
[b]Institut Universitaire de France 103, bd Saint-Michel 75005 Paris, France
[c]Université Stendhal, GIPSA-lab, Dpt. Parole et Cognition (CNRS UMR 5216)
BP25 - 38040 Grenoble Cedex 9, France
Courriel : mathilde.fort@upmf-grenoble.fr, elsa.spinelli@upmf-grenoble.fr, sonia.kandel@upmf-grenoble.fr

ABSTRACT

This study investigates the role of visual information in lexical access. We conducted a syllabic priming paradigm. Participants had to perform a lexical decision task on an acoustic target. The target was always preceded by a syllabic prime which could be displayed in audio-visual (AV), auditive (A), or visual (V) modality. The analyses on target words indicate a facilitatory priming effect for for the AV, the A and the V primes. Therefore, our results may suggest that visual information alone may activate lexical representations in the mental lexicon.

1. INTRODUCTION

Le processus de reconnaissance de mots est une activité complexe qui consiste à transformer un signal de parole en une unité de sens. La majorité des modèles qui décrivent ce phénomène postulent l'existence de niveaux de traitement différents pour les phonèmes et pour les mots (e.g. [McC86] ; [Nor00]). De nombreux travaux (e.g. [Gan80]; [Cut87]; voir [Spi05] pour une revue récente sur la question) ont mis en évidence que l'information lexicale (i.e., relative aux mots) influençait le processus de traitement des phonèmes, en modalité auditive. Par exemple, [Cut87] ont décrit le fait qu'un phonème (e.g. /b/) était plus rapidement reconnu dans un mot (e.g. belle) que dans un pseudo-mot (e.g. berre). Cet « effet de supériorité du mot » témoigne de l'influence du niveau lexical sur le niveau de décision phonémique. A l'aide de différentes méthodes (« restauration phonémique », [Sam81] ; « effet Ganong », [Gan80] ; « amorçage fragmenté », e.g. [Spi01]) d'autres travaux ont également étudié le processus d'accès au lexique. En utilisant un paradigme d' « amorçage fragmenté », [Spi01] ont montré que la présentation auditive d'une syllabe en amorce (e.g. ver) facilitait le traitement ultérieur d'un mot cible (e.g. verveine). Cette facilitation n'était observée que dans le cas où l'amorce et la cible partageaient le même début (recouvrement phonologique initial, e.g. ver-verveine), mais pas lors d'un recouvrement phonologique final (e.g. veine-verveine). Ce phénomène, appelé « effet d'amorçage », suggère donc l'existence d'un niveau lexical dans lequel la représentation d'un mot (e.g. verveine) serait activée par la présentation préalable d'une syllabe au recouvrement phonologique initial (e.g. ver), mais pas lors de la présentation d'une amorce au recouvrement phonologique final (e.g. veine). Cet « effet d'amorçage » permet donc de souligner quelles caractéristiques sont nécessaires à un signal de parole (e.g. partage du début de mot) pour qu'il puisse activer des représentations du niveau lexical.

Bien qu'un grand nombre de travaux se soient penchés sur le processus de reconnaissance de mots, la majorité d'entre eux ont –à notre connaissance- étudié ce phénomène en modalité auditive. Or, il est communément admis que le fait de voir le visage de son interlocuteur permet d'augmenter l'intelligibilité des sons de parole dans un environnement bruité (e.g. [Sum54]) ou d'en contraindre la perception en situation de conflit perceptif (l'illusion McGurk, [McG76]). En conséquence, il semblerait légitime de se demander quel rôle pourrait jouer l'information visuelle dans le processus d'accès au lexique. Cependant, peu de travaux ont à la fois étudié l'influence du niveau lexical et de l'information visuelle à la perception du signal de parole. Trois se sont penchés sur cette problématique en utilisant l'effet McGurk. Ces derniers ont trouvé des résultats contradictoires. [Bra04] et [Bar08] ont mis en évidence que l'information lexicale influençait la prise en compte de l'information visuelle au sein de l'illusion McGurk, alors que [Sam98] n'y sont pas parvenus.

En conséquence, l'objectif de cette étude consiste à déterminer si l'information visuelle seule (i.e., la gestualité oro-faciale) contribue au processus d'accès au lexique. L'originalité de travail consiste à étudier cette problématique à l'aide d'un paradigme couramment utilisé pour étudier les processus d'accès au lexique en modalité acoustique: l'amorçage fragmenté intermodal. En effet, l'utilisation de cette méthode a pour avantage d'étudier cette problématique en l'absence de conflit perceptif entre les informations visuelles et auditives (effet McGurk). De plus, celle-ci permet non seulement de collecter des données relatives aux taux de réponses correctes mais également de fournir des mesures « online » (i.e., temps de réaction) sur le processus de reconnaissance de mots. Dans notre précédente étude [For], les participants avaient pour tâche de détecter un phonème (e.g. /p/) dans un mot (e.g. troupeau) ou dans un pseudo-mot (e.g. troupi) présenté en modalité auditive seule (A) ou audiovisuelle (AV) avec ou sans bruit dans le signal acoustique (sans bruit ; -9dB ; -18dB). Les résultats ont montré qu'en présence de bruit dans le

signal acoustique, l' « effet de supériorité du mot » était plus important en modalité AV qu'en modalité A seule, suggérant que l'information visuelle contribue au processus d'accès au lexique. Le but de ce travail est de déterminer si l'information visuelle *seule* permet d'activer les représentations de mots contenues dans le lexique mental. [Kim04] ont pour cela utilisé une technique d'amorçage par répétition. Les auteurs ont observé que la présentation d'un mot en amorce en modalité visuelle seule facilitait le décodage ultérieur du même mot en modalité auditive (e.g. back-back). N'ayant pas constaté de tels résultats pour les pseudo-mots (e.g. scay-scay) les auteurs en ont conclu que l'articulation du mot « back » en amorce aurait activé la représentation lexicale du mot « back ».

Dans cette étude, nous avons décidé d'étudier cette problématique en utilisant un paradigme d'amorçage fragmenté intermodal, avec une tâche de décision lexicale. Contrairement à [Kim04], l'amorce pouvait être présentée non seulement en modalité visuelle (A), mais également en modalité auditive (A) ou en modalité audiovisuelle (AV), afin de comparer les différents effets d'amorçages obtenus en fonction de la modalité de présentation de l'amorce. Cette dernière était immédiatement suivie de la cible, toujours présentée en modalité auditive. Conformément aux résultats de [Spi01], nous devrions obtenir un effet d'amorçage facilitateur pour la condition A, lorsque la cible est un mot. Conformément aux résultats de [Kim04], nous devrions retrouver ce même résultat, pour la condition V. Enfin, pour les mêmes raisons, nous devrions également obtenir un effet d'amorçage plus important pour la condition AV que pour la condition A.

2. MÉTHODE

2.1. Participants

Soixante quatre individus (15 hommes et 49 femmes) ont participé à l'expérience. Ils étaient âgés de 17 à 38 ans (moyenne d'âge = 22.21 ans) et étaient tous de langue maternelle française. Aucun ne présentait de trouble visuel non corrigé ou auditif particulier.

2.2. Stimuli

2.2.1. Items cibles

Un corpus de 90 mots bisyllabiques (e.g. bonnet) a été sélectionné (fréquence moyenne : 24,64 occurences par millions). Chacun de ces items commençait par une syllabe de type C^1V^2 qui était toujours composée par une consonne articulée à l'avant du conduit vocal (/b/, /m/ /p/, /f/, /v/, /s/) et par une voyelle protruse (i.e., provoquant un arrondissement des lèvres: /u/, /o/, /y/). À partir de ces mots, 90 pseudo-mots bi-syllabiques ont été créés, présentant un recouvrement phonologique initial avec les mots dont ils sont issus (e.g. bonnet/bopet). Quatre-vingt dix syllabes amorces ont également été créées. Quarante-cinq d'entre elles (amorces Reliées) correspondaient à la première syllabe de la paire (e.g. bo-bonnet/bopet) alors que les quarante-cinq autres (amorces Non Reliées) n'entretenaient aucune relation (phonologique, orthographique ou sémantique) avec la cible (e.g. di-bonnet/bopet).

2.2.2. Items de remplissage

Cent quatre vingt items de remplissage (90 mots et 90 pseudo-mots) ont également été sélectionnés (e.g. di-canard/chanfi). Par conséquent, la proportion d'items Reliés étaient de 25 % seulement, afin de limiter le développement de stratégies de réponse par les participants [Slo92].

2.2.3. Enregistrements et préparation des stimuli

L'ensemble des stimuli a été enregistré en chambre sourde à l'aide d'une caméra tri-CCD (SONY DXC-990P) et d'un micro (AKG C1000S). Ils étaient prononcés par une locutrice de langue maternelle française. Ils ont été numérisés à l'aide du logiciel Dps Reality v 3.1.9 afin d'obtenir des fichiers en format *mpeg*. Le logiciel Virtual Dub a été utilisé afin d'extraire le signal acoustique et vidéo de chaque item. Chaque stimulus a également été segmenté de manière à ce que le délai entre la fin de l'amorce et le début de la cible (Intervalle Inter- Stimuli) soit de 50 ms et que le SOA (Stimulus Onset Asynchrony, i.e., le délai entre le début de l'amorce et le début de la cible) n'excède pas 250 ms [Slo92].

2.3. Procédure

Les participants étaient placés dans une pièce calme, à 50 cm d'un écran DELL (1024 x 768 pixels) et entendaient les stimuli par l'intermédiaire d'un casque SENHEISER HD 212 pro. Ils étaient instruits à l'oral qu'ils allaient percevoir une syllabe, directement suivie d'un mot ou un faux mot. Ils avaient pour tâche de décider le plus rapidement possible si la cible était un mot de la langue française en appuyant sur une des deux touches situées de part et d'autre du clavier. La main dominante était désignée pour la réponse « mot ». L'expérience se déroulait selon trois phases de 120 items chacune, variant en fonction de la modalité de présentation de l'amorce: (A), (V) ou (AV). Pour la condition V et AV, les participants étaient clairement instruits de regarder attentivement la vidéo. La présentation de l'amorce (en modalité A) ou de la cible était toujours accompagnée de l'image fixe du visage de la locutrice. L'ordre de présentation de chacune des phases et des stimuli était aléatoire. Pour chaque item cible, la condition de présentation de l'amorce ainsi que la nature du lien entre l'amorce et la cible étaient contrebalancées entre les participants. Une période d'entraînement de 10 items précédait chaque nouvelle phase.

[1] C : Consonne
[2] V : Voyelle

3.RÉSULTATS

Le taux de réponses correctes ainsi que les temps de réponses ont été calculés. Le taux d'erreurs étant très faible (< 1 %) aucune analyse n'a été effectuée sur le taux de réponses correctes. Un participant (sur 64) ainsi que 2 items cibles (sur 90) ont été écartés de l'analyse car ils présentaient des données aberrantes (Temps de réponses supérieurs à 2 écart-types de la moyenne ou supérieurs à 1500 ms, pourcentage d'erreurs supérieur à 90 %). Deux analyses de la variance (ANOVA) à mesure répétées 3 (modalité de présentation de l'amorce : A, V, AV) * 2 (lien amorce - cible : Relié vs. Non Relié) ont été effectuées par sujets (F1) et par items (F2) sur les temps de réponses.

3.1. Temps de réponse

3.1.1. Mots

Les temps de réponses sont présentés dans la figure ci-dessous (cf. Figure 1).

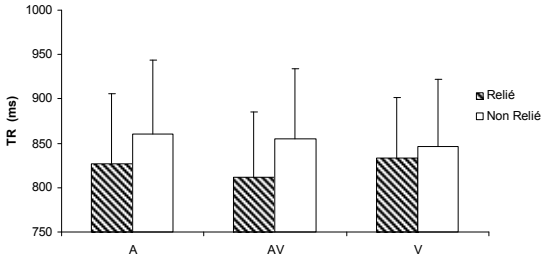

Figure 1: Temps de réponse moyens (en millisecondes, ms) en fonction de la modalité de présentation de l'amorce (A, AV ou V) et du lien entre l'amorce et la cible (Relié vs. Non Relié). Les barres verticales représentent l'écart-type pour chacune des conditions.

L'analyse statistique sur les temps de réponse (TR) révèle un effet principal du lien entre l'amorce et la cible (effet d'amorçage), *F1* (1, 62) = 64.96, p < .001, *F2* (1, 87) = 71.08, p < .001, indiquant que les cibles présentées dans la condition « Reliée » étaient plus rapidement traitées (TR moyen = 824 ms) que dans la condition « Non Reliée » (TR moyen = 854 ms). Aucun effet principal de la modalité de présentation de l'amorce n'a été obtenu, *F1* (1, 62) = 1.01, p > .05, *F2* (1, 87) = 2.15, p > .05.

Une interaction entre les deux facteurs a cependant été obtenue, *F1* (1, 62) = 4.3, p < .05, *F2* (1, 87) = 4.23, p < .05., suggérant que l'effet d'amorçage était plus important pour la condition A que pour la condition V, *F1* (1, 62) = 5.59, p < .05, *F2* (1, 87) = 3.31, p = .07, pour la condition AV par rapport à la condition V, *F1* (1, 62) = 6.75, p < .05, *F2* (1, 87) = 8.56, p < .005, mais pas pour la condition AV par rapport à la condition A, *F1* (1, 62) < 1. Les comparaisons par paires indiquent cependant qu'un effet d'amorçage est obtenu pour chacune des trois modalités. En A (TR Relié = 827 ms ; TR Non Relié = 860 ms), *F1* (1, 62) = 28.59, p < .001, *F2* (1, 87) = 19.55, p < .001, également pour la condition AV (TR Relié = 812 ms ; TR Non Relié = 850 ms), *F1* (1, 62) = 29.99, p < .001, *F2* (1, 87) = 46.07, p < .001, et de manière fondamentale pour notre étude, en modalité V (TR Relié = 833 ms ; TR Non Relié = 846 ms), *F1* (1, 62) = 3.93, p = .05, *F2* (1, 87) = 4.27, p < .05.

3.1.1. Pseudo-mots

L'analyse statistique sur les pseudo-mots ne montre aucun effet d'amorçage ni de la modalité de présentation, ni même d'interaction entre ces 2 facteurs, tous les *F1* (1, 63) étant inférieurs à 1.

4. DISCUSSION

L'objectif de notre étude était de montrer que les informations visuelles *seules* permettent d'activer les représentations lexicales. Pour cela, nous avons utilisé un paradigme d'amorçage fragmenté intermodal, avec une tâche de décision lexicale. Nous avons mesuré si la présentation d'une amorce dans différentes modalités (A, AV et V) facilitait le décodage ultérieur d'un mot ou d'un pseudo-mot cible. Les résultats suggèrent notamment que la présentation d'une syllabe en modalité visuelle seule (e.g. bo) permet d'activer les représentations lexicales partageant le même début (e.g. bonnet) facilitant par la suite la reconnaissance de ce même mot.

Un effet d'amorçage sur les mots cibles a été mis en évidence pour chacune des modalités de présentation de l'amorce. L'effet d'amorçage n'étant pas présent pour les pseudo-mots, il est possible (cf. discussion ci-dessous) que cette facilitation soit due à une pré-activation des unités lexicales par la présentation d'une syllabe pour la condition Reliée (bo-bonnet), par rapport à la condition Non Reliée (di -bonnet). En conséquence, ces résultats répliquent ceux obtenus par [Spi01] (condition A) et [Kim04] (condition V). Ils permettent donc de suggérer que les informations visuelles seules concernant la gestualité labiale de notre interlocuteur permettent d'accéder aux représentations lexicales. Or, les processus décisionnels impliqués pour répondre « oui » dans le cas où la cible était un mot et « non » dans le cas contraire ne sont pas identiques. Par conséquent, il n'est pas possible ici de comparer directement les temps de réponses obtenus pour les mots avec ceux obtenus pour les pseudo-mots. D'autres travaux sont donc en cours afin de déterminer si la facilitation observée pour les mots est due à une influence des informations visuelles sur le niveau lexical ou sur des processus de plus bas niveau (e.g. phonologique, articulatoire).

Cette étude met donc en évidence que l'information visuelle contribue au processus d'accès au lexique non seulement lorsque l'information auditive est dégradée ([For]) mais également lorsque celle-ci n'est pas disponible dans le signal de parole. Ceci suggère donc que les modèles décrivant le phénomène de reconnaissance de mots (e.g. [McC86] ; [Nor00]) devraient inclure l'information visuelle dans leur architecture. D'autres travaux sont en cours afin de

déterminer à quelle étape intervient l'information visuelle dans le processus d'accès au lexique.

L'interaction obtenue entre la modalité de présentation et le lien amorce-mot cible indique que l'effet d'amorçage est plus important pour la modalité AV et A que V seule. Ce résultat pourrait s'expliquer d'une part par le fait que l'amorce était congruente avec la cible seulement dans 25 % des cas. Il n'était donc pas pertinent -dans cette étude- de prendre en compte l'information visuelle dans les 75 % des cas restants, celle-ci étant susceptible d'induire non pas une facilitation mais une interférence avec le traitement ultérieur à effectuer sur la cible. Bien que les participants étaient clairement instruits de regarder la vidéo, il est possible que le décodage des informations visuelles soit moins efficace au fur et à mesure des essais dans cette condition [Dix80]. D'autre part, si la présentation d'une amorce en modalité visuelle seule (e.g. bo), est succeptible d'engendrer l'activation des unités lexicales partageant le même début (e.g. bonnet), elle peut également, de par sa nature, venir activer les représentations lexicales partageant le même visème (e.g. moto, poulet). En conséquence, la compétition lexicale serait donc plus accrue dans cette condition et pourrait atténuer l'effet d'amorçage dans cette condition par rapport à ceux observés en modalité auditive ou audiovisuelle.

Le fait que l'effet d'amorçage en AV ne soit pas significativement plus important que celui obtenu en A peut s'expliquer par le fait que lorsque les conditions de perception de la parole sont optimales (e.g. en l'absence de bruit [Sum54], de conflit perceptif, [McG76]), l'information auditive est suffisante pour accéder efficacement au lexique ([Spi05]). D'autres recherches doivent être menées afin de déterminer si l'information visuelle est utilisée dans toutes les situations de perception audiovisuelle de la parole ou seulement lorsque l'information auditive est indisponible ou altérée.

5.REMERCIEMENTS

Nous voudrions remercier les relecteurs anonymes qui, de par leurs critiques constructives et pertinentes, ont participé à l'amélioration de ce travail.

6. RÉFÉRENCES

[Bar08] Baratchu A., Crewther S., Kiely P. & Murphy M. (2008) "When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audio-visual speech perception", *European Journal of Cognitive Psychology*, Vol. 20(1), pp. 1-11.

[Bra04] Brancazio, L. (2004), "Lexical influences in audiovisual speech perception", *Journal of Experimental Psychology and Human Perception and Performances,* Vol. 30(3), pp. 445-463.

[Cut87] Cutler A., Mehler J., Norris D. & Segui J. (1987), "Phoneme identification and the lexicon", *Cognitive Psychology,* Vol. 19, pp. 141-177.

[Dix80] Dixon NF., & Spitz L. (1980), "The detection of auditory visual desynchrony", *Perception,* Vol. 9, pp. 719-721.

[For] Fort M., Spinelli E., Savariaux C. & Kandel S. (soumis), "The word superiority effect in audiovisual speech perception", *Speech Communication.*

[Gan80] Ganong, W. F., 3rd. (1980), "Phonetic categorization in auditory word perception", *Journal of Experimental Psychology and Human Perception and Performances,* Vol. 6(1), pp. 110-125.

[Kim04] Kim J., Davis C. & Krins P. (2004), "Amodal processing of visual speech as revealed by priming", *Cognition,* Vol. 93, pp. B39-B47.

[McG76] McGurk, H., & MacDonald, J. (1976), "Hearing lips and seeing voices", *Nature,*Vol. 264, pp. 746-748.

[Sam81] Samuel, A.G. (1981), "Phonemic restoration: Insights from a new methodology", *Journal of Eperimental Psychology: General,* Vol. 110, pp. 474-494.

[Sam98] Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998), "McGurk effect in Finnish syllabes, isolated words, and words in sentences: Effects of word meaning and sentence context", *Speech Communication,* Vol. 26, pp. 75-87.

[Sum54] Sumby, W. H., & Pollack, I. (1954), "Visual contribution to speech intelligibility in noise", *Journal of the Acoustic Society of America,*Vol. 26, pp. 212-215.

[Slo92] Slowiaczek, L.M., & Hamburger, M. (1992), "Prelexical facilitation and lexical interference in auditory word recognition" *Journal of Experimental Psychology: Learning, Memory and Cognition,* Vol. 13, pp. 64-75.

[Spi01] Spinelli, E., Segui, J., Radeau, M. (2001), "Phonological priming in spoken word recognition with bisyllabic targets", *Language & Cognitive Processes*, Vol. 16, pp. 367-392.

[Spi05] Spinelli, E. & Ferrand L. (2005), *Psychologie du langage: l'écrit, le parlé, du signal à la signification*, Armand Colin, Paris.

[McC86] McClelland, J.L., Elman, J.L. (2005), "The TRACE model of speech perception", *Cognitive Psychology*, Vol. 18, pp. 1-86.

[Nor00] Norris, D., McQueen, J.M., Cutler A., (2005), "Merging information in speech recognition: Feedback is never necessary", *Behavioral and Brain Sciences*, Vol. 23, pp. 299-325.