L'Année psychologique, 2002, *102*, 725-746

,

Note méthodologique

Université de Bourgogne, CNRS [1, 2]*
*INSERM U334, SHFJ/CEA, Orsay***
*FNRS, Université libre de Bruxelles, Belgique****

VOCOLEX : UNE BASE DE DONNÉES LEXICALES SUR LES SIMILARITÉS PHONOLOGIQUES ENTRE LES MOTS FRANÇAIS

Sophie DUFOUR[*], Ronald PEEREMAN[*], Christophe PALLIER[**], Monique RADEAU[***]

SUMMARY : VoCoLex: A lexical database on phonological similarity between French words

Several studies on auditory word recognition indicate that word processing is influenced by phonological similarity with other words. We describe a lexical database, VoColex, which provides several statistical indexes of phonological similarity between French words. Phonological similarity is computed according to two distinct principles. According to the first principle, phonologically similar words share initial phonemes with the target word. According to the second principle, phonological neighbours correspond to any words which can be derived from the target by a single phoneme change (substitution, addition, or deletion) whatever the position of the modified phoneme. The statistical data provided by VoCoLex allow the control and the empirical manipulation of various measures of phonological similarity, as well as quantitative descriptions of the auditory lexicon.

Key words : *Auditory word recognition, phonological neighborhood, cohort, lexical database.*

1. LEAD, 6, boulevard Gabriel, 21000 Dijon.
2. E-mail : Sophie@leadserv.u-bourgogne.fr

L'étude des processus cognitifs de traitement du langage profite des progrès technologiques, et en particulier informatiques. Outre la souplesse que ces progrès permettent dans l'élaboration des situations expérimentales, ils sont aussi source de stimulation de la recherche en psycholinguistique grâce à la possibilité de réaliser des descriptions quantitatives des langues. Les descripteurs quantitatifs organisés dans des bases de données lexicales permettent ensuite de contraindre les théories psycholinguistiques ainsi que la sélection et le contrôle des stimuli dans l'élaboration des situations expérimentales (Frauenfelder, Content et Peereman, 1996). L'objectif de ce texte est de décrire une nouvelle base de données lexicales, VoCoLex, portant sur diverses caractérisations des similarités phonologiques entre les mots de la langue française.

Les bases de données lexicales ont une longue tradition dans la recherche psycholinguistique. Ainsi en est-il des recueils de fréquences d'occurrence des mots dans les langues anglaise (Kucera et Francis, 1967; Carroll, Davies et Richman, 1971) et française (Imbs, 1971), ainsi que des bases de données « généralistes » plus récentes telles que la base MRC (Coltheart, 1981) ou Celex (Baayen, Piepenbrock et Gulikers, 1995) pour la langue anglaise, et Brulex (Content, Mousty et Radeau, 1990) ou Lexique (New, Pallier, Ferrand et Matos, 2001) pour la langue française. Toutefois, en dépit de l'intérêt manifeste de telles bases de données, la spécificité croissante des divers domaines de recherche en psycholinguistique a conduit au développement de nouveaux outils répondant aux besoins d'un domaine d'étude particulier. Par exemple, pour la langue française, Alario et Ferrand (1999) ont rapporté un ensemble de descripteurs pertinents pour les études portant sur la production du langage, et Peereman et Content (1999) ont développé la base de données LexOP fournissant plusieurs caractérisations des relations entre formes orthographiques et phonologiques pertinentes pour l'étude des processus de lecture (Peereman, Content et Bonin, 1998) et d'écriture (Bonin, Peereman et Fayol, 2001). La base de données VoCoLex s'inscrit dans ce courant de développement d'outils spécialisés et concerne plus particulièrement la recherche portant sur la reconnaissance des mots parlés.

La recherche empirique indique que la reconnaissance des mots parlés est influencée par la similarité phonologique entre le mot entendu et d'autres mots connus de l'auditeur. D'une

manière générale, les modèles conçoivent l'identification d'un mot comme un processus de discrimination entre des multiples candidats lexicaux (Luce, Pisoni et Goldinger, 1990; McClelland et Elman, 1986; Marslen-Wilson, 1987; Marslen-Wilson et Welsh, 1978; Norris, 1994). L'activation conjointe de mots phonologiquement similaires est suggérée par exemple, par les effets de densité et de fréquence du voisinage lexical (Frauenfelder, Baayen, Hellwig et Schreuder, 1993 ; Goldinger, Luce et Pisoni, 1989 ; Luce et Pisoni, 1998 ; Luce, Pisoni et Goldinger, 1990). Un enjeu majeur de ces études est que les effets de similarité phonologique éclairent sur la manière dont l'information lexicale est accédée et organisée. Par ailleurs, l'observation d'effets de voisinage phonologique dans les études antérieures contraint le chercheur à contrôler la similitude phonologique entre les stimuli auditifs et l'ensemble des représentations lexicales. Une telle pratique méthodologique est courante dans le domaine de la recherche portant sur la reconnaissance des mots écrits. Afin de fournir une description la plus complète possible des similitudes phonologiques entre les mots, les deux définitions actuellement envisagées (Luce, Pisoni et Goldinger, 1990 ; Marslen-Wilson et Welsh, 1978) sont ici développées: l'une en référence au modèle de la Cohorte (Marslen-Wilson et Welsh, 1978), l'autre en référence au modèle NAM (Neighborhood Activation Model; Luce, Pisoni et Goldinger, 1990).

Le modèle de la Cohorte, dans sa version initiale (Marslen-Wilson et Welsh, 1978) attribue un rôle particulier au début des mots et considère le traitement de la parole comme un processus optimalement adapté à la distribution séquentielle de l'information acoustique. Plus spécifiquement, il suppose qu'un auditeur sélectionne à l'aide des deux ou trois premiers phonèmes une cohorte de candidats alignés avec ces sons initiaux. Cette phase d'activation est suivie d'une phase de réduction progressive des hypothèses lexicales: chaque phonème rentrant élimine de la cohorte tous les candidats qui ne s'apparient plus avec l'information présente dans le signal de parole. Un mot est reconnu lorsqu'il est le seul membre restant dans la cohorte. Ce point de reconnaissance correspond au point d'unicité (PU): moment à partir duquel, un mot reste l'unique candidat à être activé. Le modèle de la cohorte prédit donc qu'un mot sera reconnu plus vite lorsqu'il dévie rapidement des autres candidats lexicaux (PU précoce) que lorsque l'information entrante

reste compatible longtemps avec plusieurs entrées lexicales (PU tardif).

Plusieurs études décrivent des données indiquant que les mots à PU précoce sont reconnus plus vite que ceux à PU tardif (par ex., Marslen-Wilson, 1984; Radeau et Morais, 1990; Radeau, Mousty et Bertelson, 1989; Wingfield, Goodglass et Lindfield, 1997). Toutefois, l'hypothèse que l'effet du PU traduirait la séquentialité du processus normal d'identification des mots a été mise en question. Selon Radeau, Morais, Mousty et Bertelson (2000), l'effet du PU résulterait de processus stratégiques favorisés par un débit lent de paroles. Avec un débit de paroles normal, proche de celui d'un mot extrait d'une conversation, aucun effet du PU n'est observé. Dans le domaine de la détection de phonèmes, cependant, certaines données s'accordent quelque peu avec la notion de traitement séquentiel des mots parlés (Frauenfelder, Segui et Dijkstra, 1990 ; Pitt et Samuel, 1995). Bien que les patrons de résultats observés soient assez compliqués lorsque les positions relatives du phonème critique et du PU sont prises en compte (en particulier, chez Pitt et Samuel, 1995), les latences de détection sont plus rapides pour un phonème se situant après le PU. La similarité phonologique entre le mot cible et d'autres mots du langage pourrait donc être fonction de la similarité entre les séquences de phonèmes initiaux.

Cette opérationnalisation du voisinage phonologique contraste avec celle envisagée par NAM (Luce et Pisoni, 1998 ; Luce, Pisoni et Goldinger, 1990). À la différence du modèle de la cohorte, NAM ne tient pas compte de la directionnalité du signal de parole et aucune portion du signal acoustique n'est privilégiée. Ce modèle conçu pour rendre compte du traitement des mots monosyllabiques définit le voisinage lexical en référence à un traitement parallèle. Selon NAM, les voisins lexicaux correspondent à tous les mots qui peuvent être générés par addition, délétion, ou substitution d'un phonème, quelle que soit sa position. Contrairement à la version initiale du modèle de la cohorte, NAM suppose que la fréquence et la densité du voisinage influencent la reconnaissance. La probabilité d'identifier un mot dépend non seulement de sa fréquence mais également du nombre et de la fréquence des voisins lexicaux. Les prédictions de NAM ont été corroborées par les résultats de plusieurs études. D'une part, il ressort que l'existence de voisins lexicaux plus fré-

quents que le mot cible ralentit sa reconnaissance. D'autre part, les mots cibles possédant de nombreux voisins sont reconnus plus lentement que les mots cibles caractérisés par une densité de voisinage faible (Luce et Pisoni, 1998).

La nature des similitudes entre mots supposées pertinentes pour le traitement est intimement liée à la conceptualisation des unités de traitement des mots. Le mot étant considéré comme une séquence de phonèmes, la similarité sera estimée en référence aux phonèmes partagés entre le mot et les autres candidats lexicaux. Néanmoins, une difficulté de cette approche est que la réalisation effective des phonèmes est notablement déterminée par les phones environnants (Liberman, Cooper, Shankweiler et Studdert-Kenney, 1967) si bien que la perception d'un son de parole semble nécessiter la prise en compte d'information acoustique sur une portion plus étendue du signal. Plusieurs travaux suggèrent en effet que l'unité syllabique constitue une unité de reconnaissance de la parole (Mehler, Dommergues, Frauenfelder et Segui, 1981 ; Pallier, Sebastian-Galles, Felguera, Christophe et Mehler, 1993). Par exemple, il est plus facile de détecter que la séquence /pal/ est incluse dans un mot cible lorsqu'elle correspond à la première syllabe de ce mot (palmier) que lorsqu'elle ne correspond pas à la première syllabe (palace). Similairement, la séquence /pa/ est plus facilement détectée lorsque le mot cible débute par cette syllabe (palace *vs* palmier; Mehler *et al.*, 1981). Sur la base de ce type de résultats, Mehler et ses collègues (Dupoux et Mehler, 1990 ; Mehler, Dupoux et Segui, 1990) ont proposé que la syllabe était non seulement l'unité de décodage de la parole, mais aussi l'unité d'accès au lexique. L'unité syllabique est également considérée comme unité de production de parole par Levelt et Wheeldom (1994). Ces auteurs envisagent l'existence d'un syllabaire interne associant les spécifications phonologiques aux caractéristiques articulatoires des syllabes de la langue, la rapidité d'accès aux syllabes étant fonction de leur fréquence d'occurrence.

L'hypothèse d'un fonctionnement cognitif sensible aux unités syllabiques s'accorde bien avec les observations d'une saillance de la syllabe déjà présente chez les nouveau-nés (Bijeljac-Babic, Bertoncini et Melher, 1993), et l'exploitation de la fréquence des transitions entre syllabes dans la segmentation du flux continu de paroles chez l'enfant de 8 mois (Saffran, Aslin et Newport, 1996). Toutefois, si les observations empiriques

recueillies en français, catalan, italien et néerlandais pointent vers un rôle de l'unité syllabique dans la perception de la parole, son rôle dans les langues ayant une structure syllabique moins claire telle que l'anglais reste partiellement controversé (Bruck, Treiman et Caravolas, 1995 ; Cutler, Mehler, Norris et Segui, 1986 ; Pitt, Smith et Klein, 1998).

Il ressort que la manière de déterminer les similarités phonologiques entre mots est directement fonction d'hypothèses relatives à la fois aux unités de traitement et à la séquentialité ou parallélisme de codage. La compréhension des processus de reconnaissance des mots nécessite par conséquent d'envisager l'impact des différentes formes de similarité phonologique sur le traitement cognitif. Les observations disponibles actuellement sont partiellement ambiguës en raison de la variabilité importante du matériel linguistique manipulé. Par exemple, alors que les données conduisant à l'hypothèse d'un traitement parallèle sont obtenues à l'aide de mots monosyllabiques, les arguments en faveur d'un traitement séquentiel proviennent d'études utilisant des mots polysyllabiques. Toutefois cette distinction n'implique pas nécessairement une différence de traitement mais résulte du fait que la manipulation du PU comme indice de traitement séquentiel est difficile avec des mots monosyllabiques. Ceci provient du fait que pour l'anglais (Luce, 1986) comme pour le français (Radeau, Morais, Mousty et Bertelson, 2000), des mots courts sont souvent imbriqués dans des mots plus longs (ex. « car » dans « cartable ») si bien que le PU des mots courts est généralement localisé après leur dernier phonème. Pour les mots plus longs dont le PU est précoce, la reconnaissance pourrait avoir lieu avant la fin du mot. Des données conciliables avec cette idée sont décrites par Grosjean (1985) et Bard, Shillcock et Altman (1988) en utilisant une procédure de dévoilement progressif *(gating)* de mots extraits de conversations.

Le but de la base de données VoCoLex est de fournir un ensemble d'indicateurs statistiques relatifs aux similarités phonologiques entre mots, permettant soit le contrôle de certaines de ces variables, soit leur manipulation empirique. Les indicateurs présents dans la base correspondent d'une part à des variables dont l'influence a été observée dans des études antérieures (nombre de voisins, fréquence des voisins, point d'unicité), et d'autre part à des variables potentiellement importantes dans le traitement. Par exemple. si la svllabe est considérée comme

unité d'accès lexical, une approche du voisinage lexical tenant compte des similarités syllabiques pourrait s'avérer pertinente. La manipulation empirique de la position du phonème déviant entre le stimulus auditif et un voisin lexical devrait permettre d'examiner l'hypothèse d'un traitement séquentiel. Par exemple, si le début du mot est critique dans l'activation des voisins phonologiques, les voisins divergeant du stimulus sur un des phonèmes initiaux devraient avoir une influence moindre que les voisins divergeant sur un des phonèmes finaux. La perspective est donc double puisqu'il s'agit de mettre à la disposition des chercheurs un outil permettant à la fois de vérifier-contrôler l'influence de certaines variables, mais permettant également de nouvelles perspectives de recherche.

LES INDICES DE SIMILARITÉ PHONOLOGIQUE

Les indices de similarité phonologique sont calculés selon deux principes différents. Chaque mot apparaissant dans la base est caractérisé par son point d'unicité et par l'étendue de son voisinage phonologique estimé sur l'ensemble de la séquence phonémique. Des estimations de la similarité phonologique en fonction de la position dans la chaîne phonémique sont fournies pour chacun des deux types d'indice. Ceci permet par exemple d'examiner la réduction progressive du nombre de mots présents dans la cohorte en fonction de chaque nouveau phonème de la séquence. La figure 1 illustre la réduction moyenne de la cohorte en fonction de la progression dans la séquence de phonèmes pour les mots bisyllabiques de 4 à 7 phonèmes. Une particularité des cohortes initiales est que leurs tailles sont fonction de la longueur des mots. Les mots plus longs sont caractérisés par des cohortes initiales plus denses que les mots courts. Ainsi, les mots longs et courts n'ont généralement pas les mêmes séquences phonémiques initiales. En français, les mots longs débutent souvent par des suites de syllabes CV simples, alors que les mots plus courts -surtout lorsqu'ils sont monosyllabiques -permettent des structures syllabiques plus complexes mais également plus rares. Les descripteurs quantitatifs sont basés soit sur un dénombrement absolu des unités lexicales similaires (comptage lexical), soit sur la fréquence des mots similaires dans la langue

(comptage textuel). Enfin, les similarités phonologiques sont estimées soit indépendamment des similarités syllabiques, soit en en tenant compte.

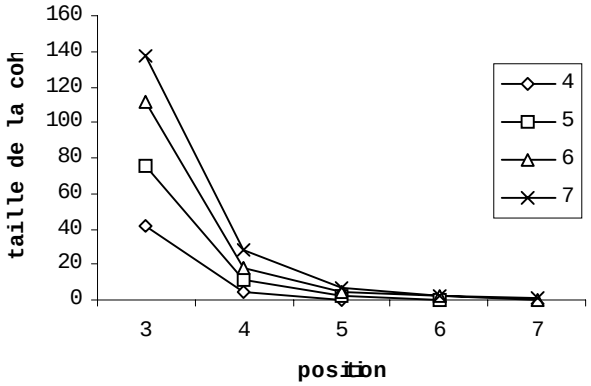

Fig. 1. - Réduction de la taille de la cohorte en fonction de la position des phonèmes pour les mots bisyllabiques de 4 à 7 phonèmes

Reduction of cohort size as a function of phoneme position for bisylabic words of 4 to 7 phonemes long

CORPUS LEXICAL

Les entrées lexicales correspondent à l'ensemble des mots de 2 à 8 phonèmes *(n* = 105464) apparaissant dans la base de données lexicales Lexique (New, Pallier, Ferrand et Matos, 2001). La limitation à 8 phonèmes répond au souci d'éviter une base de données trop volumineuse et peu maniable. Ce corpus inclut l'ensemble des mots mono- et bisyllabiques (excepté 6 entrées), ainsi que 50 893 mots trisyllabiques et 8 624 mots quadrisyllabiques. Nous pensons que ce corpus répond à la majorité des besoins expérimentaux. Les représentations phonétiques utilisées sont celles extraites de Lexique et corrigées par Peereman et Dufour (sous presse). Les seules modifications consistent en la suppression des distinctions entre les voyelles [a] et [ɑ] et entre [o] et [ɔ] .Celles-ci sont motivées par la disparition progressive de ces distinctions dans la plupart des dialectes du francais (Léon. 1992 ; Warnant, 1987). Les transcriptions pho-

nétiques sont donc basées sur 14 voyelles, 3 semi-voyelles, et 19 consonnes. Les symboles phonétiques et les codes correspondants utilisés dans la base de données VoCoLex sont similaires à ceux de Lexique (tableau I).

TABLEAU I. – *Caractères correspondant aux symboles phonétiques de VoCoLex*

Characters coding the phonetic symbols in VoCoLex

Voyelles	Exemples	Codes VoCoLex	Consonnes	Exemples	Codes VoCoLex
ɑ / a	bas, plat	a	p	père	p
e	blé	e	t	vite	t
ɛ	lait	E	k	sac	k
i	ville	i	b	robe	b
ɔ / o	mort / peau	O	d	dos	d
œ̃	brun	1	g	gare	g
u	route	u	f	fou	f
y	rue	y	s	sale	s
ø	deux	2	ʃ	chat	S
œ	peur	9	v	rêve	v
ə	le	*	z	zèbre	z
ɛ̃	train	5	ʒ	gel	Z
ɑ̃	vent	@	l	lent	l
ɔ̃	bon	§	ʀ	rose	R
Semi vowels			m	main	m
			n	nez	n
j	yeux	ï	ɲ	vigne	N
w	oui	ü	ŋ	swing	G
ɥ	huile	ÿ	x	loch	x

SYLLABIFICATION

Les syllabes non terminales des mots de la langue française sont essentiellement de structure Consonne + Voyelle et la syllabification est généralement non ambiguë. Ainsi, la césure syllabique dans des mots tels que /paʀadi/ (paradis) et /telefon/ (téléphone) se situe entre la voyelle et la consonne qui suit (/pa-Radi/, /te-le-fon/). Une telle segmentation est conforme au principe de l'attaque maximale (maximum onset principle) selon lequel les consonnes intervocaliques sont préférentiellement considérées comme attaque de la syllabe suivante pour autant que la séquence attaque + voyelle ainsi obtenue soit légale. La présence de groupements consonantiques intervocaliques rend néanmoins ambiguë la segmentation syllabique d'un grand nombre de mots français. En effet, si l'on s'accorde générale-

ment sur la non-séparation des groupements occlusive + /R/ tels que /bR/ (/abʀi/, abris) ou /pR/ (/kapʀis/, caprice), la segmentation d'un groupement tel que /st/ (/pistil/ , pistil) paraît moins claire (/pis-til/ ou /pi-stil/). Laeufer (1992) recense ainsi de nombreux groupements consonantiques français pour lesquels des propositions de segmentations différentes ont été proposées (voir tableau II).

TABLEAU II. -*Diverses propositions de syllabification de groupes consonantiques regroupées par Laeufer* (1992)

Different proposals for syllabification of consonant clusters as summarised by Lauefer (1992)

		Grammont	Delattre		Pulgram Malmberg	Noske	Levin
Groupes	Exemples		Apt.	Force			
OL	caprice	-pr	-pr	-pr	-pr	-pr	-pr
	atlas	-tl	-tl	-tl	t-l	t-l	t-l
ON	technique	-kn	-kn	-kn	k-n	k-n	k-n
OF	adverbe	-dv	-dv	-dv	d-v	d-v	d-v
OO	structure	-kt/k-t	k-t	k-t	k-t	k-t	k-t
FL	casserole	-sr	-sr	-sr	s-r	s-r	-sr
	disloque	-sl	-sl	s-l	s-l	s-l	s-l
	influent	-fl	-fl	-fl	-fl	f-l	-fl
FN	transmis	-sm	s-m	s-m	s-m	s-m	s-m
FF	blasphème	-sf/s-f	s-f	s-f	s-f	s-f	s-f
FO	diphtongue	f-t	f-t	f-t	f-t	f-t	f-t
NL	minerai	-nr	-nr	-nr	n-r	n-r	n-r
NN	calomnie	-mn/m-n	-mn	-mn	m-n	m-n	m-n
NF	hameçon	m-s	-ms	-ms	m-s	m-s	m-s
NO	samedi	m-d	m-d	-md	m-d	m-d	m-d
LL	galerie	-lr/l-r	-lr	-lr	l-r	l-r	l-r
	berlue	-rl/r-l	r-l	r-l	r-l	r-l	r-l
LN	calmant	l-m	l-m	-lm	l-m	l-m	l-m
LF	répulsif	l-s	l-s	-ls	l-s	l-s	l-s
LO	culbute	l-b	l-b	-lb	l-b	l-b	l-b

Notes : O = occlusives ; F = fricatives ; N = nasales ; L = liquides

Plusieurs études empiriques ont été menées afin d'examiner la manière dont l'auditeur syllabifie explicitement (par ex., Content, Kearns et Frauenfelder, 2001 ; Schiller, Meyer et Levelt, 1997; Treiman et Danis, 1988). Toutefois, en dépit de l'observation de tendances centrales, une variabilité interindividuelle importante dans la nature des segmentations syllabiques des mêmes séquences est également apparente. Une question supplémentaire consiste à déterminer si les segmentations syllabiques explicites imposées par des tâches telles que l'inversion

des syllabes d'un mot sont révélatrices de la manière dont les mots sont syllabifiés dans l'écoute normale de la parole. En l'absence de critère décisif en faveur d'une syllabification plutôt qu'une autre, nous avons retenu celle adoptée par Pulgram (1970). La syllabification est basée sur le principe d'une segmentation syllabique entre deux consonnes sauf dans les cas des occlusives + liquides (ex. /bʀ/, /pl/) ou d'une fricative labiodentale suivie d'une liquide *(*/fl/, /fʀ/, /vl/, /vʀ/). Une exception à cette règle est la segmentation des groupes occlusives apicodentales (/t/ et /d/) + suivies du phonème /l/. Une raison motivant la segmentation de tels groupes consonantiques (/tl/, /dl/) est qu'ils n'apparaissent jamais en début de mots dans la langue française.

PRINCIPES GÉNÉRAUX DE CALCULS

Pour chacun des mots inclus dans VoCoLex, un ensemble d'estimations quantitatives sont réalisées en référence à l'ensemble des mots rencontrés dans la base de données Lexique. Les calculs sont réalisés soit à partir des fréquences lexicales *(type),* soit en fonction des fréquences textuelles *(token).* Par exemple, le voisinage phonologique d'un mot est estimé en comptant le nombre de mots similaires, ou en sommant les fréquences d'usage des mots similaires. L' absence, du moins à notre connaissance, de bases de données fréquentielles sur les mots français parlés nous a conduit à exploiter les fréquences des mots pour le langage écrit. Les valeurs fréquentielles utilisées correspondent aux fréquences Frantext fournies dans Lexique arrondies à l'unité (champ « frantfreqparm »). Toutefois, les quelques études ayant comparé les fréquences subjectives de mots anglais à l'écrit et à l'oral indiquent de fortes corrélations (supérieures à .90) entre les deux estimations de fréquence (Howes, 1954 ; Shapiro, 1969). Des calculs de corrélations entre fréquences objectives à l'écrit et à l'oral réalisés sur la base de données Celex (Baayen, Piepenbrock et Gulikers, 1995) indiquent des corrélations de. 77 et .87 pour les mots anglais de 5 et 7 lettres, respectivement. Pour les divers calculs de voisinage, il a été tenu compte que les fréquences d'occurrence correspondent non pas aux formes phonétiques, mais aux formes orthographiques. Par consé-

quent, les fréquences de mots homophones sont cumulées pour tous les calculs.

L'ensemble des champs informatifs de VoCoLex est fourni en annexe. Les calculs portant sur la cohorte ont été réalisés à partir du second phonème de chaque mot jusqu'au phonème correspondant au point d'unicité du mot, c'est-à-dire le moment où le mot reste le seul candidat compatible avec le signal. Des estimations similaires sont aussi fournies en fonction de deux critères additionnels. Selon le premier critère, seuls les membres de la cohorte plus fréquents que le mot cible sont comptabilisés. Plusieurs données suggèrent en effet que la reconnaissance des mots est affectée par l'existence de candidats lexicaux plus fréquents que le mot cible (à l'écrit: Grainger et Segui, 1990 ; à l'oral: Luce et Pisoni, 1998; Luce, Pisoni et Goldinger, 1990). Selon le second critère, l'inclusion de candidats lexicaux dans la cohorte n'est réalisée que lorsqu'ils ont la première syllabe en commun avec le mot cible. La figure 2 illustre la réduction de la taille de la cohorte pour les mots polysyllabiques de 5 phonèmes débutant par une séquence CVC et dont la première syllabe est soit CVC, soit CV. Il apparaît que la prise en compte de la structure de la première syllabe a surtout comme effet de réduire la taille des cohortes initiales. Par contre, pour les cohortes plus tardives (définies sur les 4^{e} ou 5^{e} phonèmes) l'inclusion du critère de similarité de structure de la syllabe initiale n'a pas d'influence sur le nombre des membres de la cohorte. Il en est de même pour la position du PU qui est en moyenne de 5 que la structure syllabique soit ou non considérée. Il est probable que ceci résulte du fait que les mots différant entre eux par un phonème tardif ont généralement une structure syllabique identique *(e.g.,* /baʀaʒ/ et /baʀak/ possèdent une première syllabe CV). Ainsi, dans le cadre de l'hypothèse d'une compétition entre candidats lexicaux partageant les premiers phonèmes avec le mot cible, la prise en compte de la structure syllabique ne devrait avoir aucune influence sur la facilité de reconnaissance des mots. Toutefois, ainsi que l'indique la figure 2, la structure syllabique initiale des mots à un impact important sur la taille des cohortes initiales surtout pour les mots débutant par une syllabe CVC. La différence observée entre les mots possédant une syllabe initiale CVC ou CV résulte probablement de la structure syllabique CV dominante de la langue française. Identifier la struc-

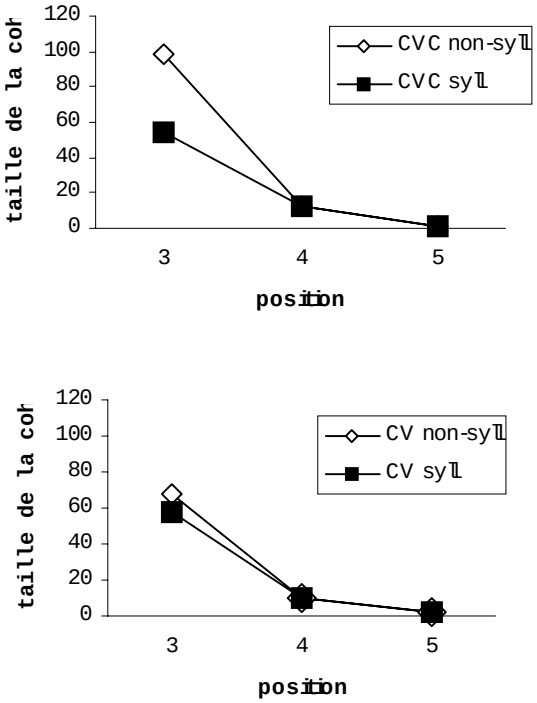

Fig. 2. - Réduction de la taille de la cohorte en fonction de la prise en compte (syll) ou non (non-syll) de la structure de la première syllabe pour les mots polysyllabiques de 5 phonèmes débutant par une syllabe CVC (graphique du haut) ou CV (graphique du bas).

Reduction of cohort size considering (*syll*) *or not* (*non-syll*) *the structure of the first syllabe for polysyllabic words of 5 phonemes long beginning with a CVC (higher graph) or CV syllabe (lower graph).*

ture syllabique initiale CVC d'un mot est donc avantageux puisque ceci conduit à éliminer un grand nombre de candidats lexicaux débutant par une syllabe CV. Inversement, identifier la première syllabe d'un mot comme étant de structure CV a peu d'influence sur la taille des cohortes initiales. Remarquons que la différence entre les deux types de mots illustrée dans la figure 2 pour le 3^{e} phonème (position 3) est encore plus importante pour le second phonème. Alors que l'inclusion du critère

syllabique réduit la cohorte de près de 90 % pour les mots débutant par une syllabe CVC, la réduction n'est que de 14 % pour les mots débutant par une syllabe CV. Le rôle possible de la taille des cohortes initiales sur la reconnaissance des mots ou dans des tâches de détection de phonèmes initiaux est actuellement encore imprécis. Néanmoins, des données recueillies à partir de mots écrits laissent suggérer que la fréquence des syllabes initiales des mots influence la reconnaissance (Perea et Carreiras, 1998).

La densité du voisinage phonologique est déterminée en référence aux entrées lexicales correspondantes au mot cible (ex. /aʀʒɑ̃/, «argent ») après substitution (/yʀʒɑ̃/, « urgent »), délétion (/aʀɑ̃/, « hareng »), ou addition (/aʀɑ̃ʒɑ̃/, « arrangeant ») d'un phonème. Cette délimitation du voisinage correspond à celle utilisée par Luce et Pisoni (1998). Comparativement à l'opérationnalisation habituelle du voisinage orthographique (Coltheart, Davelaar, Jonasson, Besner, 1977), le voisinage phonologique inclut donc des mots de longueurs différentes, ce qui n'est pas surprenant compte tenu de la nature séquentielle du signal acoustique. Il n'existe toutefois pas, à notre connaissance, d'étude contrastant les effets des différents types de voisinage sur la reconnaissance des mots. Sur l'ensemble des mots, les nombres moyens des trois types de voisins lexicaux sont de 4.5, 0.8 et 1.1, respectivement. Dans une première série de calculs, la position séquentielle du changement n'est pas prise en compte pour les calculs de voisinage, mais la densité du voisinage correspondant à chaque type de modification (substitution, délétion, addition) est fournie. Dans une seconde série de calculs, les nombres de voisins par substitution, délétion, ou addition sont fournis pour chacune des positions phonémiques (par ex. le nombre de voisins par substitution du 3^{e} phonème). La figure 3 décrit le nombre de voisins par substitution en fonction de la position du phonème modifié pour les mots de structures syllabiques CV/CVC (ex. « balade ») et CVC/CV (ex. « balcon »). On remarquera que le nombre moyen de voisins diminue après l'attaque de la première syllabe pour ensuite s'accroître pour l'attaque de la seconde syllabe. De tels indices peuvent s'avérer particulièrement intéressants dans les études recourant à la tâche de détection de phonèmes cibles dans des mots. Par exemple, d'un point de vue méthodologique, il paraît souhaitable que la comparaison des performances de détection de pho-

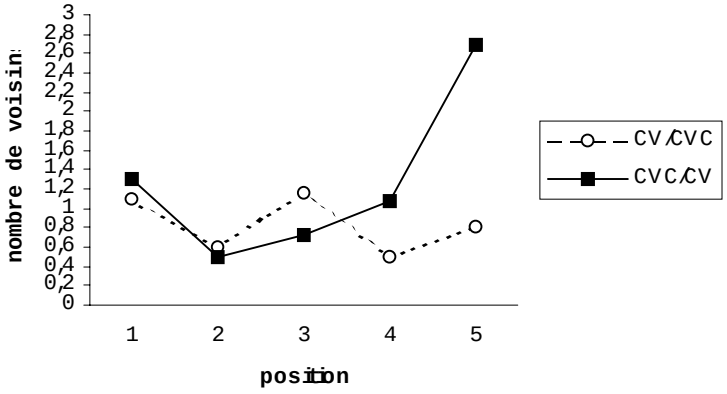

Fig. 3. - Nombre de voisins lexicaux par substitution en fonction de la position du phonème modifié pour les mots de structures CV/CVC et CVC/CV

Number of lexical substitution neighbours as a function of the position of the substituted phoneme for CV/CVC and CVC/CV words

nèmes initiaux ou finaux soit basée sur des stimuli appariés quant aux nombres de voisins différant par le phonème critique. Enfin, une estimation du voisinage est aussi réalisée en ne considérant que les mots voisins de plus haute fréquence que le mot cible qui, selon le modèle NAM, entraînerait une diminution de la performance de reconnaissance des mots.

ACCESSIBILITÉ DE LA BASE

La base de données VoCoLex (format.txt) est téléchargeable à l'adresse http://leadserv.u-bourgogne.fr/bases/vocolex. Les scripts utilisés pour les calculs de voisinages, cohortes et syllabification sont également disponibles sur internet. Ces scripts permettent donc, en fonction des besoins de l'utilisateur, de calculer les indices de similarité phonologique sur d'autres suites phonémiques, ou en référence à un autre corpus lexical. Une documentation détaillée sur l'utilisation des scripts est présentée à la même adresse internet.

ANNEXE

Noms et descriptions des champs informatifs de VoCoLex

phon : code phonétique.

hom: champ mentionnant s'il existe ou non des autres entrées homophoniques (hétérographiques ou non). Valeurs : 1 = mot homophone; 0 = mot non homophone.

graph : représentation orthographique du mot.

graphies: représentations orthographiques des entrées homophoniques.

cgram : classe grammaticale.

cgramH : classes grammaticales des homophones.

nphons : nombre de phonèmes.

Franfreqpm : fréquence formelle (selon Lexique) de l'entrée orthographique (arrondie à l'unité par million).

Frecum : fréquences cumulées des entrées orthographiques homophones (fréquence de la forme phonologique).

nsyll : nombre de syllabe.

psylpal : segmentation syllabique. Les transitions syllabiques sont représentées par le signe « -».

Calculs de cohorte

(rmq. Le mot cible n'est jamais comptabilisé dans les différentes cohortes. Il s'agit donc des nombres de mots « compétiteurs »).

PU: point d'unicité du mot.

PUs: point d'unicité du mot calculé par rapport aux mots possédant la même première syllabe.

CoPty2 à CoPty9 : nombre de candidats dans la cohorte au X^{e} phonème. La 9^{e} position correspond au silence suivant la fin du mot pour les mots de 8 phonèmes. La cohorte sur le 1^{er} phonème n'est pas fournie (elle correspond au nombre de mots ayant un premier phonème identique). **Sigles : Co = cohorte; P = phonologique ; ty = calculs par type.**

CoPto2 à CoPto9 : idem que CoPty2 à CoPty9 mais par token. Il s'agit donc de la fréquence cumulée des mots présents dans la cohorte.

HfCoPty2 à HfCoPty9 : idem que CoPty2 à CoPty9 en ne considérant que les mots plus fréquents que le mot cible. Il s'agit donc d'une fraction de la cohorte considérée dans CoPty2 à CoPty9. **Sigle: Hf = haute fréquence.**

HfCoPto2 à HfCoPto9 : idem que HfCoPty2 à HfCoPto9 mais par token. Il s'agit donc de la fréquence cumulée des mots plus fréquents dans la cohorte. **Sigle: to = calcul par token.**

SCoPty2 à ScoPty9 : nombre de candidats dans la cohorte *ayant la même première syllabe que le mot cible.* Il s'agit d'un sous-ensemble de la cohorte telle que calculée en CoPty2 à CoPty9. **Sigle: S = syllabe.**

SCoPto2 à ScoPto9 : idem que SCoPty2 à ScoPty9 mais par token. Il s'agit donc de la fréquence cumulée des mots de la cohorte ayant une même première syllabe.

HfSCoPty2 à HfSCoPty9 : idem que SCoPty2 à ScoPty9 mais en ne considérant que les mots plus fréquents.

HfSCoPto2 à HfSCoPto9 : idem que HfSCoPty2 à HfSCoPty9 par token. Il s'agit donc de la fréquence cumulée des mots de la cohorte qui sont plus fréquents que la cible, et ayant la même première syllabe.

Calculs de voisinage

Le voisinage est déterminé en prenant en compte :

- les voisins par substitution d'un caractère phonétique ;
- les voisins par addition d'un caractère ;
- les voisins par délétion d'un caractère.

Les calculs déterminent :

- le nombre total de voisins ;
- le nombre total de voisins plus fréquents ;
- le nombre de voisins par substitution ;
- le nombre de voisins par addition ;
- le nombre de voisins par délétion

(tous les calculs ci-dessus sont réalisés par type et par token).

En outre, les valeurs par type sont fournies pour :

- le nombre de voisins par substitution pour chaque position ;
- le nombre de voisins par addition pour chaque position ;
- le nombre de voisins par délétion pour chaque position.

Les différents champs informatifs sont les suivants :

- voty: nombre de voisins total par type (somme des trois sortes de voisins). **Sigle: vo = voisin** ;
- voto : idem par token ;
- vohfty : nombre de voisins total plus fréquents par type ;
- vohfto : idem par token ;
- voSty : nombre de voisins par substitution, par type. **Sigle: S = substitution** ;
- voSto : idem par token ;
- voAty : idem pour les voisins par addition. **Sigle: A = addition** ;
- voAto : idem par token ;
- voDty : idem pour les voisins par délétion. **Sigle: D = délétion** ;
- voDto : idem par token ;
- voStyl à voSty8: nombre de voisins par substitution pour chaque position, par type ;
- voAtyl à voAty8 : idem pour les voisins par addition ;
- voDty 1 à voDty8 : idem pour les voisins par délétion ;

— voHStyl à voHSty8: nombre de voisins par substitution plus fréquents pour chaque position, par type. **Sigle: H = voisin de plus Haute fréquence** ;
— voHAtyl à voHAty8 : idem pour les voisins par addition ;
— voHDtyl à voHDty8 : idem pour les voisins par délétion.

RÉSUMÉ

L'étude psycholinguistique de la reconnaissance des mots parlés indique que les similarités phonologiques du mot avec d'autres mots de la langue influencent son traitement. Nous décrivons une nouvelle base de données lexicales informatisée, VoCoLex, qui fournit un ensemble d'indicateurs statistiques sur les similarités phonologiques entre mots de la langue française. Les similarités phonologiques sont estimées selon deux principes différents. Le premier tient compte de l'ordre séquentiel des phonèmes. Selon le second, les voisins phonologiques correspondent à tous les mots pouvant être dérivés par modification d'un phonème, quelle que soit sa position. Par exemple, VoCoLex fournit le nombre de voisins phonologiques par substitution d'un phonème, ou bien encore le nombre d'items partageant les mêmes « n » premiers phonèmes, en tenant compte ou non de la structure syllabique. Ces données permettent le contrôle ou la manipulation de diverses mesures de similarités, ainsi que des descriptions quantitatives du lexique parlé.

Mots-clés : *reconnaissance des mots parlés, voisinage phonologique, cohorte, base lexicale.*

BIBLIOGRAPHIE

Alario F. X., Ferrand L. - (1999) A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition, *Behavior Research Methods, Instruments and Computers, 31*,531-552.

Baayen R. H., Piepenbrock R., Gulikers L. - (1995) *The CELEX Lexical Database*, Philadelphia (PA), Linguistic Data Consortium, University of Pennsylvania,

Bard E. G., Shillcock R. C., Altmann G. T. M. - (1988) The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context, *Perception and Psychophysics, 44*, 395-408.

Bijeljac-Babic R., Bertoncini J., Mehler J. - (1993) How do four-day-old infants categorize multi-syllabic utterances ?, *Developmental Psychology, 29*, 711-721.

Bonin P., Peereman R., Fayol M. - (2001) Do phonological codes constrain the selection of orthographic codes in written picture naming ? , *Journal of Memory and Language, 45*, 688-720

Bruck M., Treiman R., Caravolas M. - (1995) Role of the syllable in the processing of spoken English : Evidence from a nonword comparison task, *Journal of Experimental Psychology* : *Human Perception and Performance, 21*,469-479.

Carroll J. B., Davies P., Richman B. - (1971) *The American heritage word frequency book*, New York, Houghton Mifflin.

Coltheart M. - (1981) The MRC psycholinguistic database, *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Coltheart M., Davelaar E., Jonasson J. T., Besner D. - (1977) Access to the internal lexicon, *in* S. Dornic (Edit.), *Attention and Performance*, New York, Academic Press, 535-555.

Content A., Kearns R. K., Frauenfelder U. H. - (2001) Boundaries versus onsets in syllabic segmentation, *Journal of Memory and Language, 45*, 177-199.

Content A., Mousty P., Radeau M. - (1990) Brulex : une base de données lexicales informatisée pour le français écrit et parlé, *L'Année psychologique, 90*, 551-566.

Cutler A., Mehler J., Norris D., Segui J. - (1986) The syllable's differing role in the segmentation of French and English, *Journal of Memory and Language, 25*, 385-400.

Dupoux E., Mehler J. - (1990) Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code, *Journal of Memory and Language*, 29, 316-335.

Frauenfelder U. H., Baayen R. H., Hellwig F. M., Schreuder R. - (1993) Neighborhood density and frequency across languages and modalities, *Journal of Memory and Language, 32*, 781-804.

Frauenfelder U. H., Content A., Peereman R. - (1996) L'utilisation de bases de données lexicales en psycholinguistique: Applications à la reconnaissance des mots, *Actes* « *Lexique et communication parlée* », GDR-PRC, Toulouse.

Frauenfelder U. H., Segui J., Dijkstra T. - (1990) Lexical effects in phonemic processing : Facilitatory or inhibitory ?, *Journal of Experimental Psychology* : *Human Perception and Performance, 16*, 77-91.

Goldinger S. D., Luce P. A., Pisoni D. B. - (1989) Priming lexical neighbors of spoken words : Effects of competition and inhibition, *Journal of Memory and Language, 28*, 501-518.

Grainger J., Segui J. - (1990) Neighborhood frequency effects in visual word recognition.: A comparison of lexical decision and masked identification latencies, *Perception and Psychophysics, 47*, 191-198.

Grosjean F. - (1985) The recognition of words after their acoustic offset: Evidence and implications, *Perception et Psychophysics, 38*, 299-310.

Howes D. - (1954) On the interpretation of word frequency as a variable affecting speed of recognition, *Journal of Experimental Psychology, 48*, 106-112.

Imbs P. - (1971) *Études statistiques sur le vocabulaire français, Dictionnaire des fréquences, Vocabulaire littéraire des XIX^e^ et XX^e^siècles*, Centre de Recherche pour un trésor de la langue française (CNRS), Nancy, Paris, Librairie Marcel Didier.

Kucera H., Francis W. N. - (1967) *Computational analysis of present-day American English*, Providence (RI), Brown University Press.

Laueufer C. - (1992) Syllabification and resyllabification in French, *in* J. Benjamins Pub. Co, *Theoretical analyses in romance linguistics*, Amsterdam, 18-36.

Léon P. - (1992) *Phonétisme et prononciations dufrançais*, Paris, Nathan.

Levelt W. J. M., Wheeldom L. -(1994). Do speakers have access to a mental syllabary ? , *Cognition, 50*, 239-269.

Liberman A. M., Cooper F. S., Shankweiler D. P., Studdert-Kennedy M. - (1967) Perception of the speech code, *Psychological Review, 74*, 431-461.

Luce P. A. - (1986) A computational analysis of uniqueness points in auditory word recognition, *Perception and Psychophysics, 39*, 155-158.

Luce P. A., Pisoni D. B. - (1998) Recognizing spoken words : The neighborhood activation model, *Ear and Hearing*, 1-36.

Luce P. A., Pisoni D. B., Goldinger S. D. - (1990) Similarity neighborhoods of spoken words, *in* G. T. M. Altmann (Edit.), *Cognitive models of speech processing* : *Psycholinguistic and computational perspectives*, Cambridge, MIT Press, 122-147.

Marslen-Wilson W. D. - (1984) Function and process in spoken word recognition, *in* H. Bouma et D. G. Bouwhuis (Edit.), *Attention and Performance* X: *Control of language processes*, Hillsdale (NJ), Erlbaum, 125-150.

Marslen-Wilson W. D. - (1987) Functional parallelism in spoken word-recognition, *Cognition, 25*, 71-102.

Marslen-Wilson W. D., Welsh A. - (1978) Processing interactions and lexical access during word recognition in continuous speech, *Cognitive Psychology, 10*, 29-63.

McClelland J. L., Elman J. L. - (1986) The Trace model of speech perception, *Cognitive Psychology, 18*, 1-86.

Mehler J., Dommergues J. Y., Frauenfelder U., Segui J. - (1981) The syllable's role in speech segmentation, *Journal of Verbal Learning and Verbal Behavior, 20*, 298-305.

Mehler J., Dupoux E., Segui J. - (1990) Constraining Models of Lexical Access : The onset of word recognition, *in* G. T. M. Altmann (Edit.), *Cognitive models of speech processing*: *Psycholinguistic and computational perspectives*, Cambridge, MIT Press, 236-262.

New B., Pallier C., Ferrand L., Matos R. - (2001) Une base de données lexicales du français contemporain sur Internet: Lexique, *L'Année psychologique, 101*, 447-462, www.lexique.org.

Norris D. - (1994) Shortlist : A connectionist model of continuous speech recognition, *Cognition, 52*, 189-234.

Pallier C., Sebastian-Galles N ., Felguera T ., Christophe A., Mehler J. - (1993) Attentional allocation within the syllable structure of spoken words, *Journal of Memory and Language, 32*, 373-389.

Peereman R., Content A. - (1999) Lexop : A lexical database providing orthography-phonology statistics for French monosyllabic words, *Behavior Research Methods, Instruments and Computers, 31*, 376-379, ftp://ftp .ulb.ac. be/pub/packages/psyling/Lexop/.

Peereman R., Content A., Bonin P. - (1998) Is perception a two-way street ? The case of feedback consistency in visual word recognition, *Journal of Memory and Language, 39*, 151-174.

Peereman R., Dufour S. - (sous presse) Un correctif aux notations phonétiques de la base de données Lexique, *L'Année psychologique*. http://leadserv.u-bourgogne.fr/bases/lexiquecorr/.

Perea M., Carreiras M. - (1998) Effects of syllable frequency and syllable neighborhood frequency in visual word recognition, *Journal of Experimental Psychology* : *Human Perception and Performance, 24*, 134-144.

Pitt M. A., Samuel A. G. - (1995) Lexical and sublexical feedback in auditory word recognition, *Cognitive Psychology, 29*, 149-188.

Pitt M. A., Smith K. L., Klein J. M. - (1998) Syllabic effects in word processing : Evidence from the structural induction paradigm, *Journal of Experimental Psychology* : *Human Perception and Performance, 24*, 1596-1611.

Pulgrame- (1970) *Syllabe, Word, Nexus, Cursus*, The Hague, Mouton.

Radeau M., Morais J. - (1990) The uniqueness point effect in the shadowing of spoken words, *Speech Communication*, 9, 155-164.

Radeau M., Mousty P., Bertelson P. - (1989) The effect of the uniqueness point in spoken- word recognition, *Psychological Research, 51*, 123-128.

Radeau M., Morais J ., Mousty P ., Bertelson P. - (2000) The effect of speaking rate on the role of the uniqueness point in spoken word recognition, *Journal of Memory and Language, 42*, 406-422.

Saffran J. R., Aslin R. N., Newport E. L. - (1996) Statistical learning by 8-month-olds, *Science, 274*, 1926-1928.

Schiller N. O., Meyer A. S., Levelt W. J. M. - (1997) The syllabic structure of spoken words : Evidence from the syllabification of intervocalic consonants, *Language and Speech, 40*, 103-140.

Shapiro B. J. - (1969) The subjective estimation of relative word frequency, *Journal of Verbal Learning and Verbal Behavior, 8*, 248-251.

Treiman R., Danis C. - (1988) Syllabification of intervocalic consonants, *Journal of Memory and Language*, 27, 87-104.

Wamant L. - (1987) *Dictionnaire de la prononciation française*, Paris, Duculot.

Wingfield A., Goodglass H., Lindfield K. C. - (1997) Word recognition from acoustic onsets and acoustic offsets : Effects of cohort size and syllabic stress, *Applied Psycholinguistics, 18*, 85-100.

(Accepté le 1er février 2002.)