

# INFLUENCE DES FRÉQUENCES LEXICALES DES LANGUES FRANÇAISE ET DREHU SUR L'ACQUISITION DES CONSONNES INITIALES DE MOTS

Julia Monnin<sup>1,2</sup> et Hélène Lævenbruck<sup>1</sup>

<sup>1</sup>EA Transcultures, Université de la Nouvelle-Calédonie, Nouméa, France; <sup>2</sup>ICP, Département Parole et Cognition, GIPSA-lab, Grenoble, France; monninjulia@yahoo.fr; helene.loevenbruck@gipsa-lab.inpg.fr  
<http://www.gipsa-lab.inpg.fr>

## ABSTRACT

This study extends a cross-linguistic collaboration on phonological development, which aims to compare production of word-initial obstruents across sets of languages which have comparable consonants that differ in overall frequency or in the frequency with which they occur in analogous sound sequences. By comparing across languages, the influence of language-specific distributional patterns on consonant mastery can be disentangled from the effects of more general phonetic constraints on development. To extend the comparison to French, we counted type frequencies in French databases and did a preliminary experiment with French-acquiring two-year-old children. In preparation for studying the effects of Drehu exposition in Drehu and French bilingual children, we counted frequencies in Drehu.

**Keywords:** Phonological development, lexical frequency, obstruents, French, universal, language-specific, Drehu language.

## 1. INTRODUCTION

Au cours de l'acquisition phonologique, deux tendances sont observées. D'une part, certains sons dits « difficiles » sont acquis plus tardivement que d'autres qui apparaissent dès le babillage dans toutes les communautés linguistiques. Ainsi, les fricatives et les affriquées sont maîtrisées plus tard que les obstruantes [1][2]. Les consonnes dorsales sont maîtrisées plus tardivement quand elles sont suivies de /i/ que de /u/ [3]. D'autre part, ces sons « phonétiquement difficiles » sont acquis plus ou moins tôt selon les langues. Par exemple, /v/ semble être maîtrisé plus tôt en finnois, estonien et bulgare qu'en anglais. Ingram [3] suggère que la fréquence de /v/ en anglais est plus faible, ce qui explique les différences de calendrier des acquisitions. Certaines tendances de l'acquisition phonologique s'expliqueraient ainsi par des différences de phonèmes et de leurs fréquences selon les langues. Cette étude s'inscrit dans le cadre du projet Paidologos (<http://www.ling.osu.edu/~edwards/>), qui examine les influences qu'auraient les fréquences des consonnes à l'initiale des mots sur l'acquisition phonologique, dans plusieurs langues. Il s'agit de comparer les productions de consonnes d'attaque dans différentes langues présentant des distributions fréquentielles peu similaires, chez des enfants de 2 à 5

ans. L'étude présentée ici étend ce projet au français et au drehu. Il s'agit dans un premier temps de calculer des données fréquentielles pour ces deux langues, puis de présenter les premiers résultats obtenus dans un test préliminaire de répétition de mots auprès d'enfants francophones. Une étude similaire de répétition de mots en drehu est prévue.

## 2. DONNÉES FRÉQUENTIELLES DU FRANÇAIS

Certaines études suggèrent que les données fréquentielles concernant les phonèmes ou les séquences de phonèmes peuvent varier selon le registre étudié [5]. Les fréquences calculées sur des corpus de parole adressée à l'adulte peuvent ainsi différer de celles de la parole adressée à l'enfant. Cependant, le nombre de corpus spécifiquement adressés aux enfants est limité en français. Ainsi, nous avons choisi de comparer 3 différents types de données : un large corpus écrit, un vaste corpus oral adressé à l'adulte et enfin un corpus plus restreint de parole adressée au jeune enfant.

### 2.1. Corpus

Le premier corpus est le lexique adulte LEXIQUE 2 (<http://www.lexique.org>) qui a été obtenu à partir de 3200 textes écrits du français. LEXIQUE 2 contient 31 millions d'items, à partir desquels une liste de 130 000 items distincts orthographiquement a été tirée. Les transcriptions phonétiques sont disponibles, de même que les classes grammaticales. Le second corpus, LEXIQUE 3, est constitué de données orales obtenues à partir de sous-titres de films français. Il s'agit donc de dialogues parlés. LEXIQUE 3 contient 14,7 millions d'items. Les transcriptions phonétiques et les classes grammaticales sont également disponibles. Le troisième corpus est un recueil d'enregistrements effectués auprès de parents – le plus souvent la mère – s'adressant à l'enfant. Le premier recueil effectué concerne 5 enfants âgés de 22 à 26 mois. La durée des enregistrements totalise 4 heures 45 minutes et comprend 9620 mots de contenu. Les enregistrements ont été effectués à Nouméa (Nouvelle Calédonie), dans des milieux francophones monolingues, dont le français est très proche (voire identique) du français hexagonal. Ils ont été transcrits phonétiquement par une phonéticienne entraînée. D'autres enregistrements ont été ajoutés à ce corpus. Ils répondent aux mêmes critères (âge de l'enfant,

adultes s'adressant à l'enfant) et ont été extraits du corpus York [6]. Deux sessions d'enregistrements de 2000 et 3000 mots de contenus ont été sélectionnées. L'enfant avait entre 22 et 23 mois et les participants étaient le père, la mère et l'expérimentatrice. Seules les transcriptions orthographiques étaient disponibles, un phonétiseur automatique a été utilisé pour obtenir les transcriptions phonétiques [7], ensuite vérifiées et corrigées au besoin. Les données issues de nos propres enregistrements et celles du corpus YORK ont été regroupées afin de constituer notre base de données adressées à l'enfant (CDS : child directed speech).

## 2.2. Analyse des corpus

Les mots de fonctions sont souvent réduits en français. Aussi, seuls les mots de contenu et les mots de fonction accentués ont-ils été retenus dans cette étude. Les déterminants et les prépositions non accentués ont été ôtés. Tout mot répété n'a été comptabilisé qu'une seule fois (fréquence « de type »). Les phénomènes de liaisons et d'enchaînements, fréquents en français et amenant des resyllabifications ont été pris en compte. Notamment, les séquences « l'ours », « un ours », « des ours » ont été transcrits /lurs/, /nurs/ et /zurs/. De récents travaux suggèrent en effet que ces différents exemplaires sont mémorisés individuellement par l'enfant [8]. Nous avons donc considéré les versions resyllabées comme autant d'occurrences commençant par une consonne différente.

Avant de choisir quelles séquences Consonne Voyelle (CV) comparer en français avec les autres langues, nous avons répertorié 16 catégories consonantiques à l'initiale des mots, détaillées ci-dessous (symboles WorldBet) : /p, b, t, d, k, g, f, v, s, z, S (=ʃ), Z (=ʒ), l, r, m, n/.

Certaines voyelles ont été regroupées. Il s'agit des voyelles qui sont souvent substituées l'une à l'autre. Les voyelles nasales ont été comptées avec leurs homologues orales. Nous totalisons ainsi 7 catégories vocaliques :

- |                     |                           |
|---------------------|---------------------------|
| A : /a/, /ɑ/ et /ã/ | O : /o/, /ɔ/ et /õ/       |
| ø (oe) : /ø/ et /œ/ | E : /e/, /ɛ/, /ẽ/ et /œ̃/ |
| u : /u/;            | i : /i/;                  |
|                     | y : /y/                   |

Pour calculer les différentes fréquences de type relatives au français, nous avons créé des scripts *awk* permettant de comptabiliser toutes les séquences CV en début de mots. Nous avons divisé ensuite ce nombre par le nombre total de mots de contenu pour nos 3 bases de données (128 891 pour LEXIQUE 2, 137 385 pour LEXIQUE 3 et 1 108 pour les corpus CDS). Les figures 1 et 2 montrent les logarithmes des fréquences de type dans les 3 corpus, pour les consonnes et les voyelles, pour les séquences CV à l'initiale des mots de contenu.

Les fréquences obtenues sous LEXIQUE 2 indiquent que les six consonnes initiales les plus fréquentes sont /r/, /d/, /k/, /s/, /p/, /m/. Les plus rares sont /z/, /n/, /Z/, /S/ et /g/. On observe qu'à l'initiale des mots, /k/ est bien plus fréquent que /t/ ou /p/, et que /p/ est plus fréquent que /t/. Les contextes vocaliques les plus fréquents sont /E/ puis /A/. Les contextes les moins fréquents sont /u/ et /y/.

Avec LEXIQUE 3, les six consonnes les plus fréquentes en position initiale de mots sont les mêmes que celles obtenues avec LEXIQUE 2, dans un ordre légèrement différent : /d/, /k/, /s/, /r/, /m/, /p/. Les 5 consonnes les moins fréquentes sont également retrouvées : /z/, /Z/, /n/, /S/, /g/. Ici encore, /k/ est plus fréquent que /t/ et /p/, et /p/ reste plus fréquent que /t/. Les contextes vocaliques les plus fréquents sont /E/ puis /A/. Les moins fréquents sont /ø/ (oe), /y/ et /u/.

Enfin, les fréquences calculées à partir des données CDS indiquent que les consonnes les plus fréquentes sont /p/, /k/, /m/, /d/, /t/, /s/, /r/. Nous trouvons à nouveau que /k/ est plus fréquent que /t/ mais /p/ est à la fois plus fréquent que /k/ et /t/. Les cinq contextes consonantiques les moins fréquents sont un peu différents des bases de données adultes : /g/, /Z/, /z/, /v/, /S/. Le fait que /z/ soit plus fréquent dans cette base de données orale CDS s'explique probablement par les resyllabifications dues aux liaisons prises en compte lors des comptages fréquentiels. Les contextes vocaliques les plus fréquents sont /A/ puis /E/. Le contexte le moins fréquent est /y/.

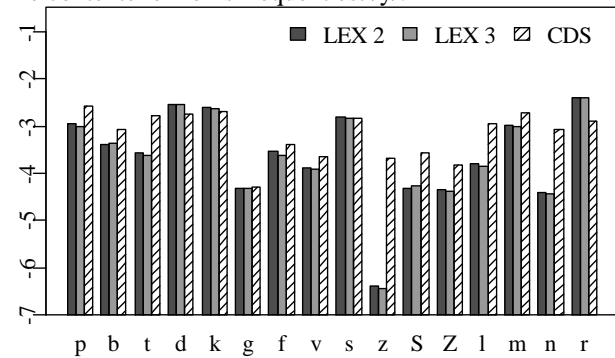


Figure 1 : Fréquence des consonnes à l'initiale de mots en français

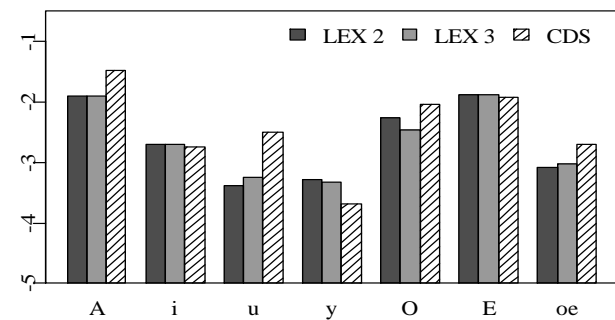


Figure 2 : Fréquence des voyelles dans les séquences initiales CV en français

Toutes nos bases de données rendent compte de résultats similaires, notamment /r/, /s/, /d/, /p/, /m/ et /k/ sont très fréquents. Une des différences concerne la fréquence de /t/, moins importante dans les bases de données adultes et assez fréquente dans la base de données adressée à l'enfant. Dans toutes les bases de données, /g/, /Z/, /z/, /S/ sont rares ; /y/ est un contexte vocalique peu fréquent alors que /E/ et /A/ sont les plus fréquents (notons que ces deux catégories représentent plus d'une voyelle).

Les fréquences relatives de /t/ et /k/ dans les séquences CV en position initiale sont différentes selon la langue

envisagée. En japonais par exemple, /k/ est bien plus fréquent que /t/ quel que soit le type de corpus [9]. Cependant en anglais, si /k/ est plus fréquent que /t/ dans les bases de données adultes, /t/ est plus fréquent que /k/ dans les données CDS [9]. Dans nos trois bases de données, /k/ est toujours plus fréquent que /t/. Dans les bases de données adultes, /k/ est plus fréquent que /p/, mais /p/ est plus fréquent que /k/ et /t/ dans le CDS.

### 3. DONNÉES FRÉQUENTIELLES DU DREHU

La langue drehu, parlée en Nouvelle Calédonie (majoritairement sur l'île de Drehu ou Lifou) compte environ 15 000 locuteurs. Elle possède en outre un riche répertoire consonantique. La situation bilingue drehu et français reste très présente parmi les enfants locuteurs du drehu. La langue drehu est parlée majoritairement aux enfants dans un contexte familial. La scolarisation se fait en français avec possibilité de quelques heures d'enseignement du drehu sans caractère d'obligation. Le système phonologique du drehu est le suivant [10]

		labio-vélaires	labiales	dentales	alvéolaires	rétroflexes	palatales	vélaires	laryngale
occlusives	orales		p	t (t)		ʈ (ʈ)	c	k	
	sonores		b	d (d)		ɖ (ɖ)	j (j)	g	
	nasales		ɱ (hm)	ɳ (hn)			ɟ (hny)	ŋ (hng)	
	sonores		m	n			ɲ (ny)	ŋ (ng)	
continues	sonores	w (g)	f	θ (th)	s			x	h
	sonores	w	v	ð (j)	z				
latérales	sonores				ʎ (ht)				
	sonores				l				
vibrante					r				

Antérieures	Postérieures	
	non arrondies	arrondies
i		u
e		o
ɛ (è)	ʌ (ò)	
a		

Quatre mères s'adressant à leur enfant en drehu ont été enregistrées à domicile. Le temps d'enregistrement totalise 2 heures 3 minutes. Les transcriptions phonétiques ont été réalisées par un phonéticien locuteur du drehu. Les données fréquentielles ont été obtenues grâce à un script awk puis divisées par le nombre total de mots de contenu (370), en suivant la même procédure que pour le français. Notre étude sur les données fréquentielles du drehu prend en compte 20 consonnes à l'initiale des mots. Nous avons établi également des regroupements vocaliques : /A/ pour /a/ et /ʌ/, /E/ pour /e/ et /è/. Nous obtenons ainsi les 5 catégories de voyelles suivantes : /A/, /i/, /E/, /u/, /O/.

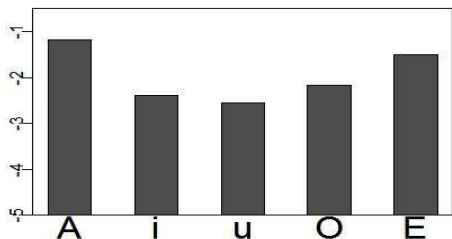


Figure 3 : Contextes vocaliques dans les séquences CV initiales du drehu (parole adressée à l'enfant)

Les résultats apparaissent dans les figures 3 et 4.

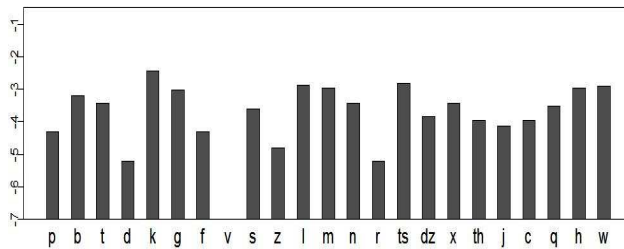


Figure 4 : Répartition fréquentielle des consonnes initiales de mots du drehu (adressé à l'enfant)

Ces figures indiquent que /k/ est plus fréquent que /t/ et /p/ à l'initiale des mots en drehu. /d/ est rare, /v/ donne une fréquence nulle. Nous retrouvons ainsi les tendances phonologiques décrites en 1983 [10] avec peu de mots contenant /v/ et /d/ en position initiale. Les contextes vocaliques /u/ et /i/ sont les moins fréquents.

## 4. ETUDE PRÉLIMINAIRE D'ENFANTS FRANÇAIS EN RÉPÉTITION DE MOTS

### 4.1. Méthodologie

Dans cette étude pilote, seules quelques séquences CV initiales ont été étudiées, afin de permettre des comparaisons avec les données déjà disponibles dans les autres langues du projet. Les consonnes /t/, /s/, et /k/ ont été analysées, dans les contextes vocaliques /a/, /i/, /y/, /u/. Les consonnes complexes /tw/ and /kw/ ont été examinées devant /a/ et /i/. Une liste de 16 mots contenant les séquences CV initiales recherchées a été créée. Les mots devaient être fréquents et ne devaient contenir qu'une à deux syllabes.

Six enfants âgés de 20 à 24 mois ont été enregistrés (une fille et cinq garçons). Ils vivaient tous à Grenoble, et tous étaient éduqués en contexte monolingue. Ils ne présentaient pas de perte auditive. Les enregistrements ont eu lieu au domicile des enfants, en présence de leur mère. Chaque essai consistait en une répétition de mot, le mot étant d'abord prononcé par l'expérimentatrice. Des images représentant les mots apparaissaient sur un écran d'ordinateur et les enfants étaient munis d'un microphone portable. Il était demandé aux enfants de répéter les mots comme ils les entendaient. Quand cela était possible, plusieurs répétitions du même mot étaient recueillies. Les réponses étaient enregistrées sur un ordinateur via CoolEdit. Pour permettre une vérification des mouvements articulatoires, un enregistrement vidéo était également effectué. Les données de deux enfants n'ont pu être étudiées, l'un refusant la tâche expérimentale et l'autre n'étant pas encore capable de produire de mot.

Une locutrice native, phonéticienne entraînée, a écouté et analysé les réponses. Les consonnes étudiées étaient notées en réponses correctes ou incorrectes. Les substitutions étaient également reportées. Une seconde phonéticienne a retranscrit une partie des données, afin de s'assurer de la fiabilité des transcriptions.

### 4.2. Résultats

La moyenne des réponses correctes pour les 4 enfants est

présentée en figure 5. /k/ et /s/ sont répétés avec moins d'erreurs que /t/ dans tous les contextes sauf /A/.

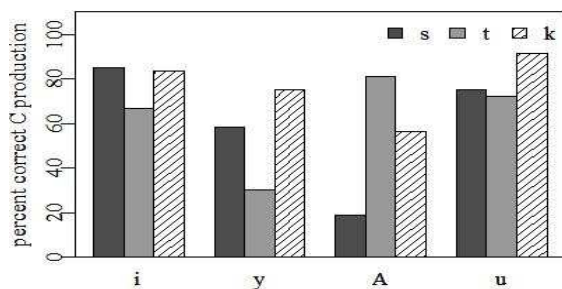


Figure 5 : Score de répétitions correctes

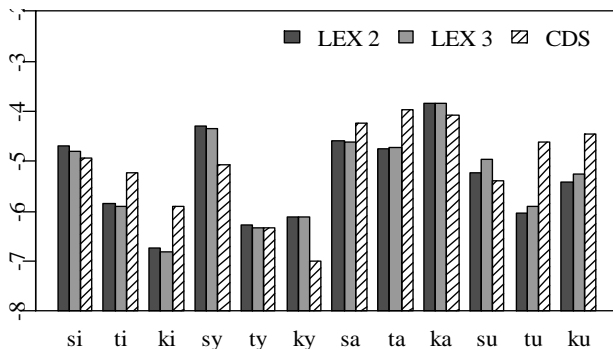


Figure 6 : Comparaison des séquences CV initiales

Une comparaison entre fréquences CV (Fig. 6) et scores de répétition correcte (Fig. 5) montre que les consonnes les moins fréquentes sont moins bien répétées.

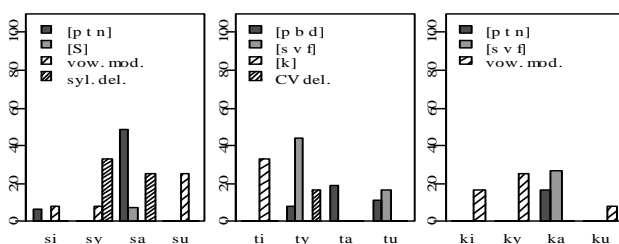


Figure 7 : Types d'erreurs en répétition de mots

Par exemple, /sa/ est moins fréquent que /ta/, et les enfants réalisent plus d'erreurs sur /s/ quand il est associé au contexte vocalique /a/. De même, /ku/ est plus fréquent que /tu/ ou /su/, et /k/ est mieux répété que /t/ et /s/ dans ce contexte. Le parallèle est observé pour /si/. Les types d'erreurs sont présentés pour chaque consonne (Fig. 7). /k/ n'est jamais remplacé par /t/, une erreur souvent faite par les enfants anglophones. Ici au contraire, /t/ est remplacé par /k/ ou des fricatives.

## 5. DISCUSSION ET CONCLUSION

Cette étude préliminaire examine le développement phonologique, chez des enfants français, des consonnes placées dans les séquences CV initiales de mots. Elle montre que /k/ et /s/ sont mieux répétés mis à part dans le contexte vocalique /A/ où /t/ est mieux réussi. De plus, /k/ n'est jamais remplacé par /t/ alors que /t/ l'est par /k/. Dans nos bases de données fréquentielles relatives au

français, /s/ et /k/ sont plus fréquents que /t/, ce qui explique probablement les meilleures réussites en répétition de mots pour ces deux consonnes. Or, /s/ est souvent considéré comme articulatoirement difficile, et /k/ fait partie des consonnes dorsales, repérées comme acquises plus tardivement [11]. Ces scores de réussite et ces types d'erreurs diffèrent de l'anglais (où /k/ est moins fréquent que /t/ dans l'input) et sont proches de ceux observés en japonais et en grec (où /k/ tend à être plus fréquent que /t/ dans les corpus de parole CDS). En drehu, /k/ est très fréquent ; /t/ est cependant légèrement plus fréquent que /s/. D'autres données doivent être recueillies afin de permettre des comparaisons plus étendues de la production de consonnes parmi différents groupes d'enfants. Cependant, ces premiers résultats semblent confirmer une influence de la langue ambiante sur le développement phonologique.

## REMERCIEMENTS

Nous remercions Fabrice Wacalie pour les transcriptions phonétiques des corpus en drehu et Mary Beckman pour ses commentaires.

## BIBLIOGRAPHIE

- [1] Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E. & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *J. of Speech and Hearing Disorders*, 55, 779-798
- [2] Hua, Z. & Dodd, B. (2000). The phonological acquisition of Putonghua (Modern Standard Chinese). *J. of Child Language*, 27, 3-42.
- [3] Davis, B.L., MacNeilage, P.F. & Matyear, C. Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, 2002, 59, 75-107.
- [4] Ingram, D. (1988). The acquisition of word-initial [v]. *Language and Speech*, 31, 77-85.
- [5] Tserdanelis, G. (2005). *The role of segmental sandhi in the parsing of speech*. Unpublished doctoral dissertation, Ohio State University, Columbus.
- [6] Plunkett, B. (2002). Null Subjects in child French interrogatives: A view from the York Corpus. In Claus D. Pusch, & Wolfgang Raible (Eds), *Romance corpus linguistics: Corpora and spoken language*, 441-452.
- [7] Bailly, G. & Alissali M. (1992). COMPOST: a server for multilingual text-to-speech system. *Traitement du Signal* 9(4), 359-366.
- [8] Chevrot, J.-P., Dugua, C. & Fayol, M. (2005). Liaison et formation des mots en français : un scénario développemental, *Langages*, 158, 38-52.
- [9] Beckman, M. E., Yoneyama, K., & Edwards, J. (2003). Language-specific and language universal aspects of lingual obstruent productions in Japanese-acquiring children. *J. of the Phonetic Society of Japan*, 7, 18-28.
- [10] Moysse-Faure, C., (1983). Le drehu, Langue de Lifou (Iles Loyauté); Paris-SELAF; 17-31
- [11] Jakobson, R. (1941/1968). Traduction : A. R. Keiler, Child language, aphasia, and phonological universals. The Hague: Mouton.