

iPhocomp : calcul automatique de l'indice de complexité phonétique de Jakielski

Hyeran Lee¹ Philippe Gambette² Melissa Barkat-Defradas^{1,3}

(1) Praxiling UMR5267 CNRS & Université Montpellier 3, Route de Mende, 34199 Montpellier

(2) LIGM, 5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée

(3) Institut des Sciences de l'Evolution UMR5554 CNRS & Université Montpellier 2,
Place Eugène Bataillon – C.C. 065 – 34095 Montpellier cedex 05

hyeranleejean@gmail.com, philippe.gambette@u-pem.fr,
melissa.barkat@univ-montp3.fr

RESUME

L'indice de complexité phonétique a été proposé par Jakielski afin d'estimer la difficulté à prononcer certains mots en anglais. Des études comparatives sur plusieurs types de pathologies du langage, dans des langues diverses, ont permis d'observer des comportements différents parmi les sujets étudiés, face à des mots d'indices de complexité phonétique variés. Nous proposons une interface web basée sur un dictionnaire phonétique afin d'estimer automatiquement l'indice de complexité phonétique d'un ensemble de mots en français : iPhocomp. Nous comparons l'étiquetage humain à l'étiquetage automatique fourni par iPhocomp afin d'estimer la performance de l'interface.

ABSTRACT

iPhocomp, automatically computing Jakielski's index of phonetic complexity

The index of phonetic complexity was introduced by Jakielski in order to estimate the difficulty to pronounce some words in English. Comparative studies about several language pathologies, in several languages, have observed distinct behavior among the studied subjects facing words with various phonetic complexity indices. We provide a web interface based on a phonetic dictionary to automatically estimate the phonetic complexity index of a set of words in French: iPhocomp. We compare the human computation of this index with the automatic computation provided by iPhocomp in order to estimate the performance of the interface.

MOTS-CLES : indice de complexité phonétique, dictionnaire phonétique, interface web.

KEYWORDS: index of phonetic complexity, phonetic dictionary, web interface.

1 Introduction

Les recherches antérieures ont montré que la complexité phonétique des mots a une influence sur leur planification et le maintien d'une production fluide de la parole, l'augmentation de la complexité nécessitant un plus grand effort pour l'activation et l'organisation des structures phonologiques et la planification motrice (Coalsen et *al.*, 2012, Howell, 2004). Contrôler cette variable est ainsi important dans les évaluations langagières ou de la mémoire utilisant des stimuli verbaux et elle peut apporter des éléments explicatifs à l'étude de la performance langagière des locuteurs.

Au cours de ces dernières années, plusieurs mesures de complexité phonétique ont été proposées : *index of phonetic complexity* (Jakielski, 2000), *phonological mean length of utterance* (Ingram et *al.*, 2001), *word complexity measure* (Stoel-Gammon, 2010), etc. Dans cette étude, nous nous intéresserons tout particulièrement à l'*index of phonetic complexity* (IPC) de Jakielski. En effet, l'IPC est une mesure complète qui permet de quantifier la complexité phonétique à travers plusieurs caractéristiques de la production langagière. Ainsi, il nous semblait intéressant d'examiner un large éventail de facteurs afin d'obtenir de plus amples informations. De plus, l'IPC a été utilisé dans divers domaines d'étude comme l'acquisition du langage (Charlier Bererd et *al.*, 2007), les pathologies du langage telle que la dyslexie (Bose et *al.*, 2011), le bégaiement (Howell et *al.*, 2006, Lasalle et *al.*, 2011), l'autisme (Ferré et *al.*, 2013), la maladie d'Alzheimer (Lee, 2012). Cet indice initialement conçu pour l'anglais a été utilisé dans différentes langues, par exemple en allemand (Dworzynski et *al.*, 2004) et en espagnol (Howell et *al.*, 2007).

L'IPC de Jakielski est basé sur la théorie de « cadre puis contenu » de MacNeilage et Davis (1990). Selon cette théorie, le développement vocal précoce est conditionné par la maturation physiologique et la maîtrise des habiletés motrices de l'appareil phonatoire, et ce de façon universelle. En effet, ces auteurs ont observé des caractéristiques régulières dans le babillage des enfants, par exemple l'apparition précoce des consonnes labiales ou coronales par rapport aux consonnes dorsales, ces dernières nécessitant un contrôle plus précis de l'appareil articuloire. De même, avec la maîtrise de la variation de la forme du conduit vocal, les consonnes fricatives émergent dans le répertoire phonétique des enfants, après les consonnes occlusives. Enfin, avec l'accroissement de la capacité articuloire, les enfants élargissent leur éventail de production des consonnes isolées aux groupes consonantiques. Ainsi, la théorie de « cadre puis contenu » possède un caractère prédictible.

Jakielski (1998) a observé 8 caractéristiques du babillage des enfants :

1. Les jeunes enfants utilisent moins fréquemment les consonnes dorsales que les enfants plus âgés.
2. Les consonnes occlusives, nasales et semi-consonnes sont récurrentes.
3. Les voyelles rhotiques sont rares.
4. Les syllabes produites ont tendance à être les syllabes ouvertes.
5. Les mots produits sont courts.
6. Les consonnes isolées sont rédupliquées dans le groupe des jeunes enfants, alors que les productions des enfants plus âgés comportent des consonnes de lieu d'articulation varié.
7. Avec l'âge, l'utilisation de groupes consonantiques augmente.

8. Les groupes consonantiques se développent de groupe consonantique homo-organique au groupe consonantique hétéro-organique.

À partir de l'observation de l'apparition précoce ou tardive de ces caractéristiques dans le babillage, Jakielski (2000) a déterminé des éléments simples ou complexes à produire et a identifié 8 paramètres afin de calculer la complexité phonétique.

Paramètres		Points non attribués	Points attribués
1	Lieu d'articulation des consonnes	Labiales, coronales, glottales	Dorsales
2	Mode d'articulation des consonnes	Occlusives, nasales, semi-consonnes	Fricatives, affriquées, liquides
3	Classe de voyelle	Monophtongues, diphtongues	Rhotiques
4	Final du mot	Se termine par une voyelle	Se termine par une consonne
5	Longueur de mot	Monosyllabique, bisyllabique	≥ 3 syllabes
6	Variation de lieu d'articulation des consonnes isolées	Rédupliquées	Variées
7	Groupe consonantique	Absence	Présence
8	Type de groupe consonantique	Homo-organique	Hétéro-organique

TABLE 1 – 8 paramètres pour calculer la complexité phonétique (Jakielski, 2000)

Dans l'IPC de Jakielski, chaque partie d'un mot composé de l'un des 8 paramètres reçoit un point de complexité. La somme de ces points permet de déterminer la valeur de la complexité d'un mot.

La valeur de complexité phonétique d'un corpus français peut être calculée à partir des règles suivantes (Eldridge, 2006) :

1. Le lieu d'articulation de la consonne : 1 point est attribué à chaque présence d'une consonne dorsale [k], [g], [ɰ], [ŋ].
Exemple : « agrafeuse » [aɡʁaføz] = 2 points.

2. Le mode d'articulation de la consonne : 1 point est attribué pour chaque consonne fricative ou liquide [f], [v], [s], [z], [ʃ], [ʒ], [l], [ʁ].
Exemple : « calculatrice » [kalkylatʁis] = 4 points.
3. La classe de voyelle : ce paramètre (i.e. présence de voyelle rhotique) n'étant pas pertinent pour le français, il a été écarté dans cette étude.
4. La fin du mot : la présence d'une consonne en position finale de mot vaut 1 point.
Exemples : « archet » [aʁʃɛ] = 0 point / « ark » [aʁk] = 1 point.
5. La longueur du mot : un nombre de syllabes supérieur ou égal à 3 correspond à 1 point.
Exemples : « anneaux » [a-no] = 0 point / « bibliothèque » [bi-bli-jo-tɛk] = 1 point.
6. La variation de lieu d'articulation des consonnes isolées : lorsque les consonnes isolées du mot appartiennent à des lieux d'articulation différents (labiales : [p], [b], [m], [f], [v] ; coronales : [t], [d], [n], [l], [s], [z], [ʃ], [ʒ], [ɲ] ; dorsales : [k], [g], [x], [ŋ]), 1 point est octroyé.
Exemple : « baguette » [bagɛt] = 1 point. Ici, on observe la variation de lieu d'articulation des consonnes isolées de labiale [b] à dorsale [g], puis de dorsale [g] à coronale [t].
En cas de présence de groupes consonantiques et d'une seule consonne isolée, ce paramètre n'est pas pris en compte.
Exemple : « empreinte » [ɑ̃pʁɛ̃t] = 0 point.
En cas de présence de groupes consonantiques et de deux consonnes isolées ou plus, la variation de lieu d'articulation est examinée entre les consonnes isolées.
Exemple : « chronomètre » [kʁɔnomɛtʁ] = 1 point. Ici, on observe la variation de lieu d'articulation de coronale [n] à labiale [m] des consonnes isolées.
7. Le groupe consonantique : 1 point est attribué à chaque présence d'un groupe consonantique.
Les consonnes produites de façon consécutive sont considérées comme un groupe consonantique. Par exemple, dans « volcan » [vɔlkɑ̃], [lk] est considéré comme un groupe consonantique, même si [l] et [k] appartiennent à des syllabes différentes.
Exemple : « rétroprojecteur » [ʁɛtʁɔpʁɔʒɛktœʁ] = 3 points.
8. Enfin, 1 point est attribué à chaque groupe consonantique hétéro-organique, c'est-à-dire impliquant des consonnes articulées en des lieux distincts.
Exemples : « agrafeuse » [agʁaføz] = 0 point. Le groupe consonantique [gʁ] est composé de deux consonnes dorsales / « brique » [bʁik] = 1 point. Le lieu d'articulation des consonnes du groupe consonantique [bʁ] varie de labiale [b] à dorsale [ʁ].

La quantification de la valeur de complexité phonétique, obtenue grâce à l'IPC de Jakielski, peut être utilisé dans divers domaines d'étude ou d'application comme l'étude de la maturation du système de la parole chez les enfants, le développement d'une méthode d'enseignement efficace dans le domaine de l'enseignement des langues étrangères en identifiant des éléments faciles ou difficiles à maîtriser, la planification d'une prise en charge orthophonique adaptée dans le domaine des pathologies du langage en discernant les éléments altérés précocement ou préservés en fonction de la complexité phonétique.

Malgré le grand intérêt qu'a suscité cette grille, il n'existait pas, jusqu'à présent, d'outil de traitement automatique de la complexité phonétique. Pourtant, automatiser le processus d'analyse réduirait le temps d'évaluation et limiterait le biais lié à la subjectivité des évaluateurs. Ainsi,

notre objectif est de développer un outil de calcul automatique de la complexité phonétique qui fournit un résultat rapide et fiable, permettant de faciliter la recherche et la mise en pratique de cette mesure.

2 Méthode

Notre méthode iPhocomp de calcul automatique de la complexité phonétique de Jakielski est décrite en Figure 1. Pour chaque mot, une transcription phonétique est obtenue à partir de la base de données Lexique 3 (New et *al.*, 2001). Les règles décrites plus haut sont alors appliquées sur cette transcription pour évaluer 8 paramètres quantitatifs. Ces paramètres sont additionnés pour obtenir la valeur de l'indice de complexité phonétique. Pour les mots dont la transcription phonétique n'a pas été trouvée dans la base de données, l'indice de complexité phonétique n'est pas évalué, mais une proposition de transcription phonétique est donnée. L'utilisateur pourra alors la corriger, et l'ajouter à un fichier personnel de transcriptions qu'il pourra utiliser en relançant le calcul.

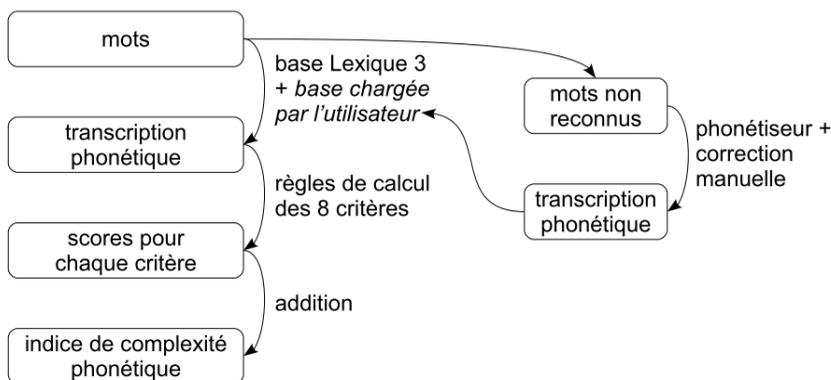


FIGURE 1 – Schéma descriptif de l'approche automatique implémentée dans iPhocomp

Cette méthode a été programmée sous forme d'une interface web à l'aide du langage PHP. Elle est disponible à l'adresse <http://igm.univ-mlv.fr/~gambette/iPhocomp/>. Le code source du logiciel est également téléchargeable à la même adresse, et réutilisable selon la licence GPLv3 afin de favoriser l'adaptation de cet outil à d'autres langues.

L'ensemble des mots à analyser est à fournir sous forme d'une simple liste de mots (un mot par ligne) dans un cadre de texte, sur la page web d'iPhocomp. L'utilisateur peut également charger une liste personnalisée de transcriptions phonétiques, qui seront prioritaires par rapport aux 125 653 transcriptions prédéfinies dans la base de données du logiciel. Ceci permet d'inclure des mots absents de notre base de données, en particulier pour des études sur les pseudo-mots, ainsi que sur les éventuelles variations de prononciation dues à des facteurs géographiques, stylistiques ou pathologiques.

L'ensemble des règles d'évaluation des 8 paramètres qui composent l'indice de complexité phonétique d'un mot, détaillées dans la Table 1, peut se déduire de sa transcription phonétique. Ainsi, nous avons choisi d'utiliser une ressource librement disponible afin de disposer de transcriptions phonétiques : la base de données Lexique 3 (New et *al.*, 2001). Lexique 3 contient

plusieurs homographes, homophones ou non. Les homographes homophones (comme « abats », qui désigne la deuxième personne du singulier du verbe « abattre » à l'indicatif, mais aussi le nom « abat » au pluriel) ont été traités comme des doublons et retirés. La base de données contient également 80 homographes non homophones, incluant par exemple les noms « actions » et « vis », qui se prononcent différemment des verbes conjugués « actions » et « vis ». Pour ces mots, seule la transcription phonétique la plus fréquente, selon la fréquence d'apparition à l'oral, également fournie par la base Lexique 3 (i.e. fréquence élaborée à partir de la transcription de dialogues de films), a été conservée. Elle a été marquée afin de pouvoir signaler ces mots comme potentiellement mal transcrits lors de l'affichage du résultat final, en couleur orange (voir « vis » en Figure 2).



Index of Phonetic Complexity

Résultats obtenus

2 transcriptions personnalisées prises en compte dans le dictionnaire phonétique.

[Cacher les résultats détaillés](#)

Mot	Prononciation	Param. 1	Param. 2	Param. 3	Param. 4	Param. 5	Param. 6	Param. 7	Param. 8	IPC
pêche	pES	0	1	0	1	0	1	0	0	3
agenda	aZ5da	0	1	0	0	1	0	0	0	2
samedi	sam [˚] di	0	1	0	0	1	1	0	0	3
vis	vis	0	2	0	1	0	1	0	0	4
jep		0	0	0	0	0	0	0	0	0
TOTAL		0	1.25	0	0.5	0.5	0.75	0	0	3

Param. 1	Param. 2	Param. 3	Param. 4	Param. 5	Param. 6	Param. 7	Param. 8	IPC
0	1.25	0	0.5	0.5	0.75	0	0	3

Attention, la liste de mots analysés contient 3 mots potentiellement problématiques : un mot non reconnu donc pas analysé, un mot avec de possibles homographes non homophones, un problème de e caduc.

Temps passé : 0.299117 ms.

4 mots reconnus / 5 : 80%.

Mots non reconnus

Les formes en minuscules sont données, la forme originale est indiquée entre parenthèses :

- jep (JEP)

Transcription des mots non reconnus

Vous pouvez, si vous le souhaitez, enregistrer une transcription phonétique des mots non reconnus pour la charger ensuite lors de l'envoi du texte, afin qu'elle soit prise en compte. Une tentative de transcription automatique basique, qui nécessite une relecture attentive, est donnée sur chaque ligne.

jep:Z[˚]p

FIGURE 2 – Interface d'iPhocomp

La base Lexique 3 propose également un traitement des « e » caducs distinct d'autres bases de données comme le TLFi (e.g. « batterie » est transcrit en [batri] selon <http://www.cnrtl.fr/definition/batterie>), en suggérant qu'ils sont prononcés, à l'aide du symbole « ° ». Les mots contenant des « e » caducs sont identifiés par iPhocomp à l'aide d'une règle simple (présence d'un « ° » dans la transcription phonétique, précédé et suivi par une seule consonne) et signalés par une couleur jaune (voir « samedi » en Figure 2).

Enfin, les mots pour lesquels il n'existe pas de transcription phonétique dans la base Lexique 3 ne sont pas analysés. Ils sont colorés en rouge et une transcription phonétique automatique basique est fournie (voir « JEP » en Figure 2). L'utilisateur est ainsi invité à vérifier soigneusement les transcriptions phonétiques des mots colorés, et éventuellement à relancer l'analyse en fournissant des transcriptions phonétiques personnalisées des mots mal transcrits.

iPhocomp fournit la valeur de complexité phonétique des mots, mais aussi les résultats détaillés pour chacun de ces 8 paramètres (Figure 2). Cela ouvre des perspectives d'analyses plus qualitatives et permet d'adapter l'IPC en fonction des corpus traités, par exemple en attribuant des poids différents aux 8 paramètres.

En plus des valeurs pour chaque mot, iPhocomp peut également calculer la valeur moyenne de la complexité phonétique d'une liste de mots. Cela permet de mener des analyses comparatives en évaluant ces divers paramètres sur deux listes de mots distinctes.

3 Résultats

Afin d'estimer la validité d'iPhocomp, nous avons procédé à une étude comparative de l'analyse manuelle et de l'analyse automatique.

Un des auteurs de cet article a effectué l'analyse manuelle d'IPC sur la base de données de Bonin et *al.* (2003). Cette base comporte un ensemble de 299 images associées à une dénomination de référence en français, ainsi que divers critères quantitatifs comme les valeurs d'accord nom-image, de familiarité conceptuelle, d'âge d'acquisition, de fréquence, de neutralité culturelle, etc. (Lee et *al.*, 2013).

Les résultats obtenus par l'analyse automatique et l'analyse manuelle ont alors été comparés. Cela a permis de corriger des erreurs initiales de conception du logiciel, par exemple sur les mots composés. Nous avons également pu repérer des erreurs dans l'analyse manuelle : sur le paramètre 2 pour « cafetière » et « scaphandrier », le paramètre 5 pour « oreiller », le paramètre 6 pour « chronomètre » et « portefeuille », et le paramètre 8 pour « flocon », « jeep » et « moustique ». Comme il n'y a pas d'ambiguïté de codage, toutes les erreurs manuelles relevaient d'erreurs d'inattention. Le calcul manuel est donc sujet à des erreurs, même s'il s'agit d'un taux très réduit (8 valeurs paramètres sur l'ensemble des 2093 valeurs de paramètres évaluées, soit moins de 0.4 % d'erreur).

Un second test a été effectué afin d'évaluer le temps d'analyse gagné sur un corpus correspondant à un cas d'usage : 118 mots extraits du discours de 13 patients atteints de la maladie d'Alzheimer, ainsi que le même nombre de mots extraits du discours d'individus sains. En utilisant la transcription phonétique fournie par Lexique 3, nous avons constaté que le temps de calcul humain des 8 paramètres se montait à environ 2 heures et demie pour ces 236 mots,

alors que le calcul automatique avec iPhocomp sur le même corpus prenait 0.54 ms. Il est à noter que si une intervention manuelle est nécessaire, en raison des types de corpus ou des choix de transcription décrits dans la partie 2, cela peut engendrer un temps de correction supplémentaire. On peut prévoir 10 secondes supplémentaires par mot à corriger puisqu'il ne s'agit pas de faire une transcription complète mais une simple modification de celle-ci.

4 Conclusion et perspectives

iPhocomp permet de réduire significativement le temps total d'évaluation de l'indice de complexité phonétique, par rapport à l'évaluation manuelle. Il permet aussi d'éviter l'erreur humaine due à l'aspect répétitif de la tâche ou à une mauvaise application des règles de calcul des paramètres, optimisant ainsi la comparabilité des résultats de différentes études.

Il nous semble important d'approfondir l'étude de la complexité phonétique de la langue française (e.g. paramètre 3). Nous pensons que l'intégration des aspects phonologiques (Pellegrino, 2009) et l'étude de divers phénomènes observables (e.g. l'ordre d'émergence des segments, la maîtrise de certaines unités, les difficultés de production, etc.) dans l'acquisition du langage, dans l'apprentissage de langues étrangères et dans les pathologies du langage en français, permettraient de dégager des éléments pertinents pour enrichir nos connaissances et notre compréhension de la complexité phonétique, permettant de mieux adapter l'indice de complexité phonétique de Jakielski aux spécificités de la langue française.

iPhocomp peut être utilisé pour toute langue disposant des ressources linguistiques adaptées. Ainsi, il nous semble intéressant d'étendre cette interface aux autres langues afin de faciliter les études translinguistiques.

Le logiciel iPhocomp est librement mis à la disposition de la communauté scientifique. Nous espérons que ce programme donnera un nouvel élan d'une part aux recherches fondamentales pour mieux appréhender les propriétés physiques et structurelles des unités sonores du langage et d'autre part aux études appliquées telles que celles conduites dans le champ de la phonétique clinique.

Références

- BONIN, P., PEERELMAN, R., MALARDIER, N., MEOT, A., & CHALARD, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, 35(1):158–167.
- BOSE, A., COLANGELO, A., & BUCHANAN, L. (2011). Effect of phonetic complexity on word reading and repetition in deep dyslexia. *Journal of Neurolinguistics*, 24:435–444.
- CHARLIER BERERD, M., & JUHEM, A. (2007). *Evolution de la complexité phonéto-phonologique et sélection lexicale chez des enfants français entre 12 et 27 mois*. Mémoire d'orthophonie, Université Lyon 1.
- COALSON, G., BYRD, C., & DAVIS, B. (2012). The influence of phonetic complexity on stuttered speech. *Clinical linguistics & phonetics*, 26(7):646–659.

- DWORZYNSKI, K., & HOWELL, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of fluency disorders*, 29:149–173.
- ELDRIDGE, K. (2006). *Phonological complexity and speech disfluency in young children*. Thèse de doctorat, University of Pittsburgh.
- FERRE, S., DOS SANTOS, C., & BONNET-BRILHAULT, F. (2013). Phonological delay or phonological impairment in autism: an intergroup comparison. *International meeting for autism research*, San Sebastian, Espagne.
- HOWELL, P. (2004). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary issues in communication science and disorders*, 31:123–140.
- HOWELL, P., & AU-YEUNG, J. (2007). Phonetic complexity and stuttering in Spanish. *Clinical linguistics & phonetics*, 21(2):111–127.
- HOWELL, P., AU-YEUNG, J., SCOTT YARUSS, J., & ELDRIDGE, K. (2006). Phonetic difficulty and stuttering in English. *Clinical linguistics & phonetics*, 20(9):703–716.
- INGRAM, D., & INGRAM, K. (2001). A whole-word approach to phonological analysis and intervention. *Language, speech, and hearing services in schools*, 32:271–283.
- JAKIELSKI, K. (1998). Motor organization in the acquisition of consonant clusters. Thèse de doctorat, University of Texas at Austin.
- JAKIELSKI, K. (2000). Quantifying phonetic complexity in words: An experimental index. *Child phonology conference*, Cedar Falls, IA.
- LASALLE, L., & WOLK, L. (2011). Stuttering, cluttering, and phonological complexity: case studies. *Journal of fluency disorders*, 36:285–289.
- LEE, H. (2012). *Langage et Maladie d'Alzheimer : Analyse multidimensionnelle d'un discours pathologique*. Thèse de doctorat, Université Montpellier 3.
- LEE, H., GAMBETTE, P., GAYRAUD, F., & BARKAT-DEFRADAS, M. (2013). Élaboration d'un outil d'évaluation des performances en dénomination pour les patients bilingues atteints de la maladie d'Alzheimer. *Rééducation orthophonique*, 253:143–152.
- MACNEILAGE, P., & DAVIS, B. (1990). Acquisition of speech production: Frames, then content. In M. Jeannerod (Ed.), *Attention and performance XIII: Motor representation and control*. Hillsdale: Lawrence Erlbaum.
- 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
- PELLEGRINO, F. (2009). *De l'identification des langues à la complexité phonologique*. Habilitation à diriger des recherches, Université Lumière Lyon 2.
- STOEL-GAMMON, C. (2010). The word complexity measure: description and application to developmental phonologic and disorders. *Clinical linguistics & phonetics*, 24:271–282.