Un correctif aux notations phonétiques de la base de données Lexique

Ronald Peereman, Sophie Dufour

R. Peereman
Sophie Dufour

Un correctif aux notations phonétiques de la base de données Lexique

In: L'année psychologique. 2003 vol. 103, n°1. pp. 103-108.

Résumé
Résumé
La base de données Lexique a été récemment développée par New, Pallier, Ferrand et Matos (2001) pour fournir de nouvelles évaluations de fréquence d'utilisation des mots écrits de la langue française. La base de données comporte également des champs informatifs additionnels facilitant la recherche en psycholinguistique. En particulier, chaque entrée orthographique est associée à une transcription phonologique. Cependant, en raison de plusieurs erreurs de conversion graphème-phonème, les codes phonologiques ne sont pas utilisables pour la sélection de stimuli ou pour la réalisation d'analyses statistiques sur le corpus de mots. Le but du manuscrit est de décrire le travail correctif qui a été effectué sur les codes phonologiques. Les modifications concernent environ 18 % de l'ensemble des entrées phonologiques de la base Lexique. Le fichier incluant les modifications est téléchargeable sur internet.
Mots-clés : base de données psycholinguistiques, Lexique, phonologie.

Abstract
Summary : A corrective to the phonetic notations of the Lexique database
The Lexique database has been recently developed by New, Pallier, Ferrand, and Matos (2001) to provide new frequency estimations of word usage. The database also includes additional information fields that can facilitate stimulus selection for psycholinguistic research in both visual and auditory word recognition. In particular, each orthographic entry is associated with a phonological transcription. However, due to several grapheme-to-phoneme conversion errors, the actual phonological codes are not useable for stimulus selection or statistical analyses on the word corpus. The purpose of the present manuscript is to describe the corrective work that has been done on the phonological codes. The computer file including the phonetic corrections can be downloaded from the web.
Key words : psycholinguistic databases, Lexical, phonology.

L'Année psychologique, 2003, *103*, 103-108

NOTES MÉTHODOLOGIQUES

LEAD / Université de Bourgogne
CNRS UMR 5022[1]

UN CORRECTIF AUX NOTATIONS PHONÉTIQUES DE LA BASE DE DONNÉES LEXIQUE

Ronald PEEREMAN[2] et Sophie DUFOUR[3]

SUMMARY : A corrective to the phonetic notations of the Lexique database

The Lexique database has been recently developed by New, Pallier, Ferrand, and Matos (2001) to provide new frequency estimations of word usage. The database also includes additional information fields that can facilitate stimulus selection for psycholinguistic research in both visual and auditory word recognition. In particular, each orthographic entry is associated with a phonological transcription. However, due to several grapheme-to-phoneme conversion errors, the actual phonological codes are not useable for stimulus selection or statistical analyses on the word corpus. The purpose of the present manuscript is to describe the corrective work that has been done on the phonological codes. The computer file including the phonetic corrections can be downloaded from the web.

Key words : psycholinguistic databases, Lexical, phonology.

Récemment, la base de données *Lexique* a été développée par New *et al.* (2001) dans le but premier de fournir des indications objectives sur la fréquence d'utilisation actuelle des mots de la langue française. Un tel outil est d'un intérêt considérable pour

1. 6, boulevard Gabriel, 21000 Dijon.
2. E-mail : peereman@u-bourgogne.fr
3. Les correctifs proposés dans ce texte sur les notations phonétiques ont été introduits dans la version actuelle de *Lexique* (http://www.lexique.org).

la recherche en psycholinguistique ainsi que d'un point de vue pédagogique. Cette entreprise s'avère d'autant plus intéressante que les tables fréquentielles dont nous disposions étaient basées sur des corpus de textes relativement anciens (Imbs, 1971). Outre les fréquences d'occurrence des mots de la langue écrite, la base *Lexique* fournit un ensemble d'informations dont les codes phonétiques (en équivalent machine des codes de l'IPA) des mots répertoriés. Cette information est intéressante pour le chercheur en psycholinguistique dans la mesure où elle permet de sélectionner du matériel expérimental également à partir de critères phonologiques. Un aspect important des codes phonétiques de *Lexique* est qu'ils ont été dérivés de manière algorithmique à partir de l'orthographe des mots.

La difficulté de la conversion automatique de chaînes orthographiques en chaînes phonétiques varie en fonction des systèmes d'écriture considérés. Les systèmes comportant de nombreuses régularités dans les correspondances entre unités orthographiques et phonologiques sont plus faciles à formaliser à l'aide de règles d'associations que les systèmes plus irréguliers. Dans les systèmes d'écriture alphabétiques décrits comme « quasi réguliers », tels que le français ou l'anglais, les procédures de conversion reposent généralement sur l'un ou l'autre des deux principes suivants. Le premier consiste à réaliser un inventaire explicite et plus ou moins précis des règles de correspondance entre unités orthographiques et phonologiques. Ces règles sont ensuite appliquées afin de coder phonologiquement le corpus de mots (par ex. Coltheart, Curtis, Atkins et Haller, 1993). Afin de permettre le codage phonologique correct des mots possédant une prononciation exceptionnelle, un lexique minimum est également requis. Le second principe consiste à recourir à des réseaux composés d'unités orthographiques et phonologiques élémentaires et d'entraîner ces réseaux à coder correctement les unités orthographiques des mots en unités phonologiques par l'intermédiaire de rétroactions correctrices. Une fois entraîné, la connaissance acquise par le réseau permet d'encoder de nouvelles chaînes orthographiques (par ex. Seidenberg et McClelland, 1989). Les données comparatives disponibles, à ce jour, suggèrent que des performances assez comparables sont obtenues par les deux principes de conversion (Coltheart, Rastle, Perry, Langdon et Ziegler, 2001 ; Plaut, McClelland, Seidenberg et Patterson, 1996).

Le système utilisé pour générer les codes phonétiques des « mots » répertoriés dans la base de données Lexique se rapproche du premier principe de conversion. Le système, baptisé LAIPTTS-F (disponible gratuitement sur internet à l'adresse http:\\www.unil.ch\imm\docs\LAIP\LAIPTTS.html) est un logiciel de synthèse sonore de la parole continue développé pour la langue française par Keller et Zellner Keller (1998) à l'Université de Lausanne. Selon la description des auteurs, le système inclut un module de traduction graphémo-phonétique de textes en phonèmes qui repose sur 540 règles de conversion. En outre, afin de pouvoir coder correctement les mots ayant une prononciation exceptionnelle, un dictionnaire comportant 7 000 mots est juxtaposé au système de règles. Dans le système complet, le module de conversion de textes en phonèmes est complété par un module prosodique et d'un module de génération du signal. Dans une étude statistique récente portant sur les formes phonologiques des mots de la langue française (Dufour, Peereman, Pallier et Radeau, 2002), nous avons désiré exploiter les codifications phonétiques générées par le module de conversion textes-phonèmes de LAIPTTS-F et présentées dans la base de données *Lexique*. L'examen des codes phonétiques a toutefois révélé plusieurs imperfections lorsque les codes étaient comparés aux notations phonétiques données par le dictionnaire *Le Petit Robert*. Signalons que si plusieurs de ces erreurs (par ex. le positionnement des schwas) sont sans conséquence dans un système destiné à la synthèse sonore de la parole (le but ultime du système LAIPTTS-F), il en est autrement pour la recherche en psycholinguistique[1]. Par exemple, la réalisation pleine d'une voyelle ou sa réduction en semi-voyelle à une influence sur la structure et la segmentation syllabique des mots. Dans une première étape, nous avons comparé les codes phonétiques de *Lexique* à ceux fournis dans la base de données psycholinguistiques *Brulex* (Content, Mousty et Radeau, 1990). L'avantage de cette comparaison réside dans le fait que les codes phonétiques des mots présents dans *Brulex* ont été introduits manuellement à partir des notations phonétiques du *Petit Robert* et non pas générés automatiquement. Il en résulte que le

1. Précisons également que certaines des codifications phonétiques fournies par le système LAIPTTS-F sont spécifiquement adaptées au couplage du système avec les bases de données diphoniques *Mbrola* (disponibles sur internet).

taux d'erreurs dans les notations phonétiques de la base *Brulex* est très faible.

Parmi les entrées orthographiques lexicales apparaissant à la fois dans *Lexique* et *Brulex* (environ 30 000), environ 2 500 différences de codifications phonétiques entre *Lexique* et *Brulex* furent dénombrées. Bien que certaines de ces erreurs correspondent à des mots dont la prononciation est exceptionnelle et non dérivable par règle, de nombreuses autres erreurs résultent de l'utilisation de règles de conversion incorrectes. Par exemple, les voyelles /i/ précédant la semi-voyelle /j/ sont généralement omises, la voyelle nasale est généralement omise pour les adverbes se terminant par « ent » ; les séquences « gia » sont codées /Zija/ au lieu de /Zja/, les mots incluant le groupe de lettres « mont » (ex. « montrer ») sont codés sans le /t/ (/m§Re/), la terminaison « ex » de certains mots n'est pas codée (/Eks/), les groupes « illi » sont codés « iji » même lorsque la prononciation « ili » s'impose..., etc. Les erreurs relevées concernent approximativement une quarantaine de problèmes de conversion. Les entrées lexicales de *Lexique* présentant ces problèmes ont toutes été corrigées en utilisant la même source pour les formes phonétiques que celle de la base de données *Brulex*. En outre, de nombreux rectificatifs ont également été apportés aux mots possédant une prononciation exceptionnelle. Soulignons que les corrections n'ont pas été apportées quand concernant les distinctions entre les voyelles /a/ (antérieure, postérieure) et /o/ (ouvert, fermé) car ces distinctions sont actuellement jugées comme disparaissant dans la plupart des dialectes du français (Léon, 1992 ; Warnant, 1987). La majorité des erreurs devrait avoir été considérée et rectifiée.

Outre ces modifications, l'ensemble des codes phonétiques de *Lexique* furent retraités pour le positionnement des schwas. Cette opération s'avéra nécessaire d'une part parce que quelques erreurs furent constatées, et d'autre part parce que des codes phonologiques homophones ne possédaient pas le (ou les) schwas en des positions similaires (par ex. pour le verbe conjugué à la 3^{e} personne du singulier ou du pluriel du présent de l'indicatif). Il semble en effet souhaitable de maintenir les principes de positionnement des schwas les plus constants possibles, ceci afin de permettre par exemple des calculs de similarité phonologique entre mots. La plupart des corrections et des homogénéisations des positionnements des schwas ont été réalisées de manière

algorithmique à partir de règles dérivées de la base de données *Brulex*. Cet algorithme fonctionne par consultation des codes phonétiques dont les schwas ont tous été supprimés, et en suggérant la présence de schwas à la suite de certaines séquences phonémiques. Dans un certain nombre de cas, la représentation orthographique du mot est également consultée afin de valider l'insertion d'un schwa ou non en fonction de la présence ou de l'absence de la lettre E. Sur l'ensemble de mots de *Brulex*, le taux de réussite de cet algorithme est proche de 100 %. Les cas non prévus par l'algorithme furent traités séparément ensuite. Cette opération sur les schwas a conduit à modifier de nombreux codes phonétiques des mots de *Lexique*. Ainsi, l'entièreté des corrections et homogénéisations des codes phonétiques porte sur environ 18 % des entrées lexicales de *Lexique*[1]. Nous pensons que l'ensemble des modifications apportées devrait permettre un contrôle plus rigoureux des caractéristiques phonologiques des mots dans les études psycholinguistiques en langue française.

L'ensemble des correctifs est disponible sur internet à l'adresse suivante : http://leadserv.u-bourgogne.fr/bases/lexiquecorr/. Les divers scripts utilisés sont également accessibles.

RÉSUMÉ

La base de données Lexique a été récemment développée par New, Pallier, Ferrand et Matos (2001) pour fournir de nouvelles évaluations de fréquence d'utilisation des mots écrits de la langue française. La base de données comporte également des champs informatifs additionnels facilitant la recherche en psycholinguistique. En particulier, chaque entrée orthographique est associée à une transcription phonologique. Cependant, en raison de plusieurs erreurs de conversion graphème-phonème, les codes phonologiques ne sont pas utilisables pour la sélection de stimuli ou pour la réalisation d'analyses statistiques sur le corpus de mots. Le but du manuscrit est de décrire le travail correctif qui a été effectué sur les codes phonologiques. Les modifications concernent environ

1. Dans l'ensemble, sachant que la longueur moyenne des entrées lexicales est de 6,8 phonèmes, les modifications apportées concernent environ 2,5 % des caractères phonétiques générés par LAIPTTS-F. Cette approximation est encore certainement grossie par le fait que plusieurs des modifications apportées n'ont pas ou peu d'incidence pour un système de synthèse de parole continue. Par exemple, en excluant les problèmes liés à la localisation des schwas, le pourcentage de caractères phonétiques modifiés se situe autour de 1,2 %. Cette estimation est très proche des estimations des erreurs décrites par Boula de Mareüil *et al.* (1998) dans un test de différents systèmes, dont LAIPTTS-F.

18 % de l'ensemble des entrées phonologiques de la base Lexique. Le fichier incluant les modifications est téléchargeable sur internet.

Mots-clés : base de données psycholinguistiques, Lexique, phonologie.

BIBLIOGRAPHIE

Boula de Mareüil P., Yvon F., d'Alessandro C., Aubergé V., Bagein M., Bailly G., Béchet F., Foukia S., Goldman J.-P., Keller E., O'Shaughnessy D., Pagel V., Sannier F., Véronis J., Zellner B. — (1998) Evaluation of grapheme-to-phoneme conversion for text-to-tpeech synthesis in French, *Proceedings of First International Conference on Language Resources & Evaluation*, Granada, Espagne, 641-645.

Coltheart M., Curtis B., Atkins P., Haller M. — (1993) Models of reading aloud : Dual-route and parallel-distributed-processing approaches, *Psychological Review, 10*, 589-608.

Coltheart M., Rastle K., Perry C., Langdon R., Ziegler J. — (2001) DRC : A dual route cascaded model of visual word recognition and reading aloud, *Psychological Review, 109*, 204-256.

Content A., Mousty P., Radeau M. — (1990) Brulex. Une base de données lexicales informatisées pour le français écrit et parlé, *L'Année Psychologique, 90*, 551-566.

Dufour S., Peereman R., Pallier C., Radeau M. — (2002) VoCoLex : une base de données lexicales sur les similarités phonologiques entre les mots français, *L'Année Psychologique, 102*, 725-746, http://leadserv.u-bourgogne.fr/bases/vocolex

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.

Keller E., Zellner Keller B. — (1998) Motivations for the prosodic predictive chain, Proceedings of ESCA *Symposium on Speech Synthesis*, Paper 76, p. 137-141, Jenolan Caves, Australie.

Léon P. — (1992) *Phonétisme et prononciations du français*, Paris, Nathan.

Mbrola, *The Mbrola Project*, Mbrola Project Development Team Faculte Polytechnique de Mons, MULTITEL-TCTS Lab, http://tcts.fpms.ac.be/synthesis/mbrola.html.

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.

Plaut D. C., McClelland J. L, Seidenberg M. S., Patterson K. E. — (1996) Understanding normal and impaired word reading : Computational principles in quasi-regular domains, *Psychological Review, 103*, 56-115.

Robert P. — (1986) Micro-Robert, *Dictionnaire du français primordial*, Paris, Dictionnaires *Le Robert*.

Seidenberg M. S., McClelland J. L. — (1989) A distributed developmental model of word recognition and naming, *Psychological Review, 96*, 523-568.

Warnant L. — (1987) *Dictionnaire de la prononciation française*, Paris, Duculot.

(Accepté le 10 janvier 2002.)