MANULEX-INFRA

Ce document reprend certaines parties du site développé par Ronald Peereman au LEAD à l'Université de Bourgogne. Visitez le site ici pour des informations plus complètes et pour télécharger d'autres ressources.

DESCRIPTION

Manulex-infra consiste en un ensemble de bases de données dérivées de *Manulex*, une base de données accessible sur internet fournissant les fréquences des mots pour 48.886 entrées lexicales rencontrées dans 54 livres scolaires en usage dans l'enseignement élémentaire. Le développement de *Manulex-infra* a été motivé par l'observation que les bases de données actuellement disponibles pour l'étude de l'acquisition de la lecture et de l'écriture présentent deux limitations majeures. Premièrement, plusieurs des bases de données linguistiques concernant la langue française sont basées sur des corpus de textes destinés aux adultes, ceci rendant leurs utilisations inappropriées pour l'étude chez les enfants. Deuxièmement, les bases de données fondées sur l'analyse de textes écrits actuels destinés aux enfants ne fournissent que des estimations statistiques relatives à la fréquence d'occurrence des mots. En dépit de leurs intérêts majeurs, la fréquence des mots n'est pas l'unique facteur influençant la performance de lecture et d'écriture, et des bases de données plus spécialisées sont donc nécessaires. En particulier, il est bien établi maintenant que les caractéristiques statistiques du langage, telle que la consistance des relations entre orthographe et phonologie, influencent l'acquisition de la lecture et de l'écriture. Ainsi, les bases de données linguistiques destinées à l'étude de la lecture et de l'écriture chez les enfants de l'école élémentaire doivent fournir des statistiques objectives sur les variables infra-lexicales, et spécifiquement concernant les relations graphème-phonème et phonème-graphème.

Les bases de données linguistiques jouent un rôle central en rassemblant un ensemble d'estimations quantitatives et objectives à propos des variables principales qui influencent l'acquisition de la lecture et de l'écriture. Nous décrivons un nouvel ensemble de bases de données sur l'orthographe Française dont la particularité principale est d'avoir été générées à partir de livres scolaires utilisés par les enfants de l'école élémentaire. Notre objectif est ainsi de compléter la base de données *Manulex* en fournissant des estimations quantitatives sur plusieurs variables infra-lexicales (syllabes, relations graphème-phonème, bigrammes...) et lexicales (voisinage lexical, homophonie, homographie). Ces analyses devraient permettre des descriptions quantitatives du langage écrit rencontré par les lecteurs débutants, la manipulation et le contrôle des variables dans les études expérimentales, et le développement de méthodes éducatives en relation avec les caractéristiques distributionnelles de l'écrit.

Toutes les entrées du lexique des formes orthographiques de *Manulex* ont été utilisées pour les calculs, exceptés les abréviations, les séquences euphoniques, les interjections, et les entrées lexicales composées (incluant un espace, une apostrophe, ou un tiret). La prononciation de nombreux noms propres apparaissant dans *Manulex* étant soit ambiguë, ou inconnue, nous n'avons conservé que les noms propres ayant une fréquence d'au moins .10 dans le lexique correspondant à l'ensemble des niveaux scolaires (niveaux 1 à 5). Le nombre total des entrées lexicales entrant dans les analyses est de 45080 pour l'ensemble des niveaux réunis (1 à 5), de 10861 pour le niveau 1 (CP), de 18131 pour le niveau 2 (CE1), et de 42422 pour les niveaux 3 à 5 (CE2 à CM2).

La plupart des calculs réalisés nécessitent un **codage phonologique** des entrées orthographiques. *Manulex* ne fournissant pas cette information, les codes phonologiques ont été ajoutés à la base de données. Les codes phonologiques ont été partiellement importés de deux bases lexicales pour le Français, BRULEX (Content et al., 1990), et LEXIQUE ver. 2 (New et al., 2004; voir aussi Peereman & Dufour, 2003, pour les corrections phonétiques incorporées à la base LEXIQUE) après conversion dans un système phonétique commun. Plusieurs des codes phonétiques ont été corrigés en accord avec la prononciation du Français

standard fournie par le dictionnaire Le Petit Robert (Version CD-Rom 2.2). Les codes correspondant aux mots absents des bases lexicales, y compris ceux correspondant aux noms propres, ont été ajoutés manuellement. Les représentations phonologiques sont basées sur 16 voyelles, 3 semi-voyelles, et 18 consonnes.

Les **codes phonétiques** correspondent à des caractères standards afin d'assurer la compatibilité entre les différents systèmes. Les caractères sont identiques à ceux utilisés dans la base lexicale LEXIQUE ver. 2. Les correspondances avec les caractères IPA sont données ci-dessous. Remarquons que le caractère dièse (#) est aussi utilisé pour définir les associations graphème-phonème afin d'indiquer un graphème silencieux (e. g., le graphème "t" à la fin du mot "fort").

IPA	codes	Exemples	IPA	codes	Exemples
Voyelles			Consonnes		
i	i	lire, vie	p	p	loupe, pain
u	u	joue, ours	t	t	terre, vite
y	y	bulle, sud	k	k	qui, bec
e	e	fée, nez	b	b	cube, brosse
ɛ	E	jouet, aile	d	d	danse, aide
a	a	date, plat	g	g	gare, bague
ɑ	A	tâche, bois	f	f	foule, phare
ø	2	deux, peu	s	s	tasse, cerf
œ	9	neuf, fleuve	ʃ	S	chat, vache
ə	*	le, ancre	v	v	vent, rêve
ɔ	o	roche, sol	z	z	zéro, rose
o	O	jaune, mot	ʒ	Z	gel, juge
ɔ̃	§	nom, pont	m	m	main, femme
ɛ̃	5	cinq, plein	n	n	nage, laine
ɑ̃	@	vent, blanc	ɲ	N	ligne, peigne
œ̃	1	un, brun	l	l	lune, pull
Semi-Voyelles			ʁ	R	rue, air
j	j	feuille, lieu			
w	w	soie, watt	ŋ	G	viking, ring
ɥ	8	huit, fruit			

La **segmentation en unités syllabiques** de chacune des entrées phonologiques est nécessaire pour estimer la fréquence des syllabes en fonction du niveau de scolarité des enfants. La segmentation syllabique du Français est généralement non ambiguë, la plupart des mots incluant des syllabes CV (consonne + voyelle). Toutefois, plusieurs mots comportent également des groupes consonantiques intervocaliques tels que /bR/ ou /st/, comme dans les mots "abris" /abRi/ et "pistil" /pistil/. La présence de groupes consonantiques intervocaliques rend parfois la segmentation syllabique ambiguë et plusieurs solutions ont été proposées (voir par exemple, Laeufer, 1992). En accord avec plusieurs travaux précédents de syllabification, les principes de segmentation proposés par Pulgram (1970) ont été adoptés. Brièvement, les frontières syllabiques sont localisées entre les consonnes adjacentes excepté lorsque des plosives ou des fricatives labio-dentales sont suivies par des liquides (e.g. /bR/, /pl/, /fl/, /fR/, /vl/, /vR/). La segmentation syllabique de chaque mot est fournie dans les bases de données.

Les entrées lexicales de *Manulex* ont été segmentées en **unités graphémiques** afin de pouvoir estimer la fréquence et la consistance des relations entre orthographe et phonologie. Autant que possible, le principe

central guidant la segmentation graphémique est que chaque groupement orthographique ne corresponde (dans la plupart des cas, voir ci-dessous) qu'à un seul phonème. Le terme "graphème" est donc utilisé ici pour désigner une lettre ou un groupe de lettres associé à un phonème. Remarquons que l'orthographe Française inclut de nombreux graphèmes multi-lettres, tels que "ou", "an", "un", "in", "eu", "ch", et "gn" .

La segmentation graphémique des mots Français n'est en général pas problématique, bien que des choix s'imposent dans plusieurs cas. Pour complexifier la tâche, à l'image d'un Rubik's cube, les solutions adoptées pour une segmentation particulière ont souvent des incidences sur plusieurs autres segmentations, et il est donc illusoire de considérer une segmentation graphémique particulière indépendamment de l'ensemble des autres segmentations graphémiques déjà effectuées. Les choix réalisés étaient gouvernés par un second principe selon lequel les segmentations devaient maximiser la mise en évidence des inconsistances de prononciations des chaînes orthographiques. Le nombre total de graphèmes obtenus est de 125.

CALCULS

Les calculs sont de deux natures différentes: caractéristiques de longueur des mots, et caractéristiques qui sont fonction du niveau de scolarité. Les caractéristiques de longueurs sont le nombre de lettres, de phonèmes, de graphèmes, et de syllabes de chaque mot. Contrairement aux caractéristiques de longueur, les autres caractéristiques dépendent du corpus analysé et donc du niveau scolaire. Les quatre corpus lexicaux de *Manulex* sont G1, G2, G3-5, et G1-5, c'est-à-dire, les mots rencontrés dans les livres scolaires de première année (CP), de seconde année (CE1), de troisième, quatrième et cinquième années, et de la première à la cinquième année. Les entrées lexicales de ces quatre corpus sont accompagnées des estimations statistiques dans les bases de données G1, G2, G3-5, et G1-5, respectivement.

Les calculs prennent en compte soit les fréquences lexicales des mots (analyses "par-type"), soit les fréquences textuelles des mots (analyses "par token"). Les calculs "par type" sont réalisés en ne prenant pas en compte la fréquence réelle des mots dans les textes. Ainsi, un mot fréquent tel que "dans" a le même poids qu'un mot plus rare tel que 'rang", ceci simplement parce que ces deux mots n'apparaissent qu'une seule fois chacun dans la liste des mots du corpus (fréquence lexicale de 1). Les calculs "par token" prennent en compte la fréquence réelle des mots dans les textes. Cette fréquence textuelle est fournie par l'index U de la base *Manulex*. Les détails des calculs sont présentés ci-dessous.

Fréquence des associations Graphème-Phonème, consistance GP et consistance PG. Les ambiguïtés dans le codage phonologique de l'écrit, et les ambiguïtés de codage orthographique du langage parlé sont généralement estimées par des index de consistance. Dans *Manulex-infra*, la consistance G-P est égale à la fréquence avec laquelle une association graphème-phonème apparaît divisée par la fréquence totale du graphème quelle que soit sa prononciation. Par exemple, la consistance GP de l'association "ch" -> /S/ (comme dans le mot "chat /Sa/) est obtenue en divisant la fréquence d'occurrence de l'association "ch" -> /S/ par la fréquence du graphème "ch" indépendamment de sa prononciation (qui peut correspondre à /S/ mais aussi à /k/ comme dans le mot "choral" /koRal/). L'index de consistance obtenu est multiplié par 100. Sa valeur maximale est donc de 100. Similairement, la consistance PG est égale à la fréquence avec laquelle une association phonème-graphème apparaît, divisée par la fréquence totale du phonème, multiplié par 100.

Les valeurs de consistance peuvent différer fortement en fonction de la position sérielle des unités graphémiques dans le mot. Ainsi, la morphologie dérivationnelle du Français fait que les finales de mots sont souvent silencieuses, et que l'orthographe est donc moins transparente. Afin de mieux caractériser les associations orthographe-phonologie, les estimations de fréquence et de consistance des associations ont été réalisées en fonction de la position relative des unités dans le mot (position initiale, positions centrales, position finale). Les résultats des calculs sont fournis pour chaque entrée lexicale, ainsi que les valeurs moyennes et les valeurs portant sur l'association la moins fréquente et consistante du mot.

Finalement, des tables distinctes fournissent les statistiques moyennes sur la consistance et la fréquence de toutes les associations rencontrées dans les corpus lexicaux.

Fréquences des unités infra-lexicales. Les fréquences des bigrammes, des biphones et des syllabes ont été calculées pour chacune des entrées lexicales des quatre corpus. La fréquence des bigrammes correspond à la fréquence avec laquelle une séquence de deux lettres cooccurrent dans la liste de mots. Transposé à la phonologie, la fréquence des biphones correspond à la fréquence avec laquelle une séquence de deux phonèmes cooccurrent dans la liste de mots. Finallement, la fréquence des syllabes est également calculée à partir de la segmentation syllabique de chacune des entrées lexicales phonologiques. Les calculs sont réalisés par type et par token, et ils tiennent compte de la position relative des unités (bigrammes, biphones, syllabes) dans le mot (positions initiale, centrale, finale). Des bases suplémentaires fournissent les statistiques moyennes sur la fréquence des bigrammes, biphones, et syllabes. Des tables sont également disponibles pour la fréquence des lettres, des phonèmes, et des trigrammes.

Voisinage orthographique et phonographique. La densité du voisinage lexical estime les similarités entre les formes lexicales. Le voisinage orthographique d'un mot est habituellement défini comme correspondant au nombre de mots qui lui sont similaires, à une lettre près (substitution d'une lettre). Ainsi, les mots LIRE, RARE, et RIME sont des voisins orthographiques du mot RIRE. La densité du voisinage lexical des mots a été estimée séparément pour les quatre corpus (niveaux 1, 2, 3-5, 1-5) puisqu'elle dépend de l'étendue du vocabulaire considéré. Les estimations sont réalisées par type et par token. Les mesures par type correspondent au nombre de voisins. Les mesures par token prennent en compte la fréquence des mots voisins en additionnant leurs fréquences. Les mots voisins qui sont homographes ne sont comptabilisés qu'une seule fois et leurs fréquences sont additionnées.

La performance de lecture à voix haute chez l'adule est facilitée par une catégorie particulière de voisins orthographiques, ceux qui sont également phonologiquement similaires au mot cible à prononcer (Peereman & Content, 1997). Ces voisins lexicaux, qualifiés de "phonographiques" sont donc à la fois orthographiquement et phonologiquement similaires au mot cible. La densité du voisinage phonographique a donc été calculée pour chaque entrée lexicale. La similarité phonologique entre les mots est déterminée similairement à la similarité orthographique, à l'exception du fait que les séquences phonologiques sont cette fois considérées. Ainsi, des mots sont phonologiquement voisins lorsqu'ils ne diffèrent que par un seul phonème (substitution d'un phonème). Le résultat des calculs sont incorporés dans les quatre bases principales et les voisins ainsi que leurs fréquences sont listés dans des fichiers séparés (NO1, NO2, NO35, NOall pour les voisins orthographiques; NOP1, NOP2, NOP35 et NOPall pour les voisins phonographiques). ici rubrique Télécharger\Bases de données

Homophones et homographes. Le nombre de mots homophones et le nombre de mots homographes ont été calculés pour chacune des entrées lexicales des quatre corpus. A nouveau, les calculs sont réalisés par type et par token. Les valeurs ont été ajoutées aux quatre bases principales (Manu1, Manu2, Manu35, ManuAll). Les mots homophones hétérographes (phonologies identiques et orthographes différentes, e. g., "FIN" et "FAIM") et les mots homographes hétérophones (orthographes identiques et phonologies différentes, e. g., "elles COUVENT" et "le COUVENT") de chacune des entrées lexicales sont listés avec leurs fréquences dans des fichiers distincts (HP1, HP2, HP35, et HPall pour les HomoPhones hétérographes; HG1, HG2, HG35, et HGall pour les HomoGraphes hétérophones). ici rubrique Télécharger\Bases de données

Point d'unicité orthographique. Dans les études portant sur la reconnaissance des mots parlés, le point d'unicité est habituellement défini comme correspondant à la position sérielle du phonème (en comptant à partir du premier phonème du mot) à partir duquel le mot devient unique, et diverge donc de tout autre candidat lexical. Transposé aux formes orthographiques, le point d'unicité orthographique réfère à la position de la lettre (en comptant de gauche à droite à partir de la première lettre du mot) à partir de laquelle le mot devient unique et diverge donc de tout autre candidat lexical. Le point d'unicité orthographique est déterminé pour chacune des entrées lexicales des quatre corpus.

SIGNIFICATION DES ABRÉVIATIONS

(A partir de la page 7, les significations des abréviations utilisant un lexique très simple ne sont plus traduites de l'anglais.)

DESCRIPTIONS DES MOTS

ORTHO : forme orthographique

PHON : forme phonologique (codes phonétiques)

SYNT : Catégorie morphosyntaxique : NC-nom, NP-nom propre, VER-verbe, ADJ-adjectif, ADV-adverbe, PRO-pronom, PRE-préposition, CON-conjonction, INT-interjection, DET-déterminant, ABR-abréviation, and UEUPH-marque euphonique

U : Fréquence d'Usage pour 1 million de mots

PSYLL : segmentation syllabique avec '.' comme séparateurs

NBSYLL : nombre de syllabes

GSEG : segmentation graphémique avec '.' comme séparateurs

PSEG : segmentation phonémique reliée à celle graphémique

GPMATCH : associations graphème-phonème reliant les champs GSEG et PSEG par des '-'séparés par des '.'. Le caractère le plus à gauche est un '(' qui indique le début du mot. Le caractère le plus à droite est un ')' qui indique un mot se terminant par. Ces deux caractères peuvent être utilisés pour trouver certaines associations graphème-phonème au début ou à la fin des mots. Par exemple, la recherche avec '(ch-S.' ou '.ch-S)' fournit la liste des mots comprenant respectivement l'association 'ch-S' au début et à la fin des mots.

nbLET : nombre de lettres

nbPHON : nombre de phonèmes

nbGRAPH : nombre de graphèmes

puortho : point d'unicité orthographique

HOMOPHONES ET HOMOGRAPHES

nbHPty : nombre d'HomoPhones, Type count

nbHGty : nombre d'HomoGraphes, Type count

nbHPNGty : nombre d'HomoPhones Non homoGraphes, Type count

nbHGNPty : nombre d'HomoGraphes Non homoPhones, Type count

nbHPto : nombre d'HomoPhones, Token count

nbHGto : nombre d'HomoGraphes, Token count

nbHPNGto : nombre d'HomoPhones Non homoGraphes, Token count

nbHGNPto : nombre d'HomoGraphes Non homoPhones, Token count

BIGRAMMES ET BIPHONES

frBIGtty : mean BIGram frequency, Type count

frBIGtto : mean BIGram frequency, Token count

frBIGity : frequency of the Initial BIGram, Type count

frBIGito : frequency of the Initial BIGram, Token count

frBIGmty : mean frequency of the middle BIGrams, Type count

frBIGmto : mean frequency of the middle BIGrams, Token count

frBIGfty : frequency of the Final BIGram, Type count

frBIGfto : frequency of the Final BIGram, Token count

frBIPtty : mean BIPhone frequency, Type count

frBIPtto : mean BIPhone frequency, Token count

frBIPity : frequency of the Initial BIPhone, Type count

frBIPito : frequency of the Initial BIPhone, Token count

frBIPmty : mean frequency of the Middle BIPhones, Type count

frBIPmto : mean frequency of the Middle BIPhones, Token count

frBIPfty : frequency of the Final BIPhones, Type count

frBIPfto : frequency of the Final BIPhones, Token count

SYLLABES

frSYLity : frequency of the Inital SYLlable, Type count

frSYLito : frequency of the Initial SYLlable, Token count

frSYLmty : mean frequency of the Middle SYLlables, Type count

frSYLmto : mean frequency of the Middle SYLlables, Token count

frSYLfty : frequency of the final SYLlable, Type count

frSYLfto : frequency of the final SYLlable, Token count

VOISINAGE LEXICAL

nbONty : number of Orthographic Neighbors, Type count

nbONto : number of Orthographic Neighbors, Token count

nbPGNty : number of PhonoGraphic Neighbors (phonological AND orthographic neighbors), Type count

nbPGNto : number of PhonoGraphic Neighbors (phonological AND orthographic neighbors), Token count

MEAN FREQUENCY AND CONSISTENCY OF GRAPHEME-PHONEME AND PHONEME-GRAPHEME ASSOCIATIONS

frGPtty : mean frequency of Grapheme-Phoneme associations, Type count

frGPtto : mean frequency of Grapheme-Phoneme associations, Token count

coGPtty : mean consistency of Grapheme-Phoneme associations, Type count

coGPtto : mean consistency of Grapheme-Phoneme associations, Token count

coPGtty : mean consistency of Phoneme-Grapheme associations, Type count

coPGtto : mean consistency of Phoneme-Grapheme associations, Token count

Minimal Frequency and Consistency of Grapheme-Phoneme and Phoneme-Grapheme Associations

frGPmity : frequency of the Grapheme-Phoneme associations having the Minimal value on Type count

frGPmito : frequency of the Grapheme-Phoneme associations having the Minimal value on Token count

coGPmity : consistency of the Grapheme-Phoneme associations having the Minimal value on Type count

coGPmito : consistency of the Grapheme-Phoneme associations having the Minimal value on Token count

coPGmity : consistency of the Grapheme-Phoneme associations having the Minimal value on Type count

coPGmito :consistency of the Grapheme-Phoneme associations having the Minimal value on Token count

Frequency and Consistency of Grapheme-Phoneme and Phoneme-Grapheme Associations by Position (initial, middle, final)

frGPity : frequency of the Inital Grapheme-Phoneme association, Type count

frGPito : frequency of the Initial Grapheme-Phoneme association, Token count

frGPmty : mean frequency of the middle Grapheme-Phoneme associations, Type count

frGPmto : mean frequency of the middle Grapheme-Phoneme associations, Token count

frGPfty : frequency of the Final Grapheme-Phoneme association, Type count

frGPfto : frequency of the Final Grapheme-Phoneme association, Token count

coGPity : consistency of the Inital Grapheme-Phoneme association, Type count

coGPito : consistency of the Initial Grapheme-Phoneme association, Token count

coGPmty : mean consistency of the middle Grapheme-Phoneme associations, Type count

coGPmto : mean consistency of the middle Grapheme-Phoneme associations, Token count

coGPfty : consistency of the Final Grapheme-Phoneme association, Type count

coGPfto : consistency of the Final Grapheme-Phoneme association, Token count

coPGity : consistency of the Inital Phoneme-Grapheme association, Type count

coPGito : consistency of the Initial Phoneme-Grapheme association, Token count

coPGmty : mean consistency of the middle Phoneme-Grapheme associations, Type count

coPGmto : mean consistency of the middle Phoneme-Grapheme associations, Token count

coPGfty : consistency of the Final Phoneme-Grapheme association, Type count

coPGfto : consistency of the Final Phoneme-Grapheme association, Token count

Notes

Les fréquences des associations phonème-graphème sont égales aux fréquences des associations graphème-phonème.

Les calculs sur les tokens sont basés sur le *U* de Manulex.