Manuel de Lexique Infra

Lexique-Infra 1.00 basée sur **Lexique 3.83**

L'objectif principal de cette base de données est de fournir des statistiques infra-lexicales pour un très grand nombre de mots polysyllabiques (plus de 130 000) provenant d'un corpus adulte. Pour les mots de Lexique 3.83, nous proposons des calculs de consistance et de régularité des associations graphème-phonème et phonème-phonème d'un mot et ceci calculé selon la position dans le mot (initiale, milieu, finale) mais aussi par fréquence de type et de token). Pour les différents fichiers constituant cette base, une description de chaque champ de la base est donnée ci-dessous. Les fréquences utilisées dans tous les calculs sont les fréquences sous-titres de Lexique.

Fichier “Lexique.Infra.Corresp.Graphème-Phonème.xlsb”

Item (Item)
Les formes orthographiques utilisées extraites de Lexique.

Phonologie (Phonologie)
La forme phonologique donnée par Lexique.

Voyelles		
Codes Lexique	**Exemples**	**Sons nommés**
a	bat, plat	A
i	lit, émis	I
y	lu	U
u	roue	Ou
o	peau, mot	o (fermé)
O	éloge, fort	o (ouvert)
e	été	e-fermé
E	paire, treize	e-ouvert
°	abordera	schwa élidable
2	deux	e-fermé
9	œuf, peur	e-ouvert
5	cinq, linge	in (voy. Nasale)
1	un, parfum	un (voy. nasale)
@	ange	an (voy. nasale)
§	on, savon	on (voy. nasale)
3	parvenu	schwa non élidable
Semi-Voyelles		
j	yeux, paille	y (semi-voyelle)
8	huit, lui	ui (semi-voyelle)
w	oui, nouer	w (semi-voyelle)

Consonnes		
Codes Lexique	**Exemples**	**Sons nommés**
p	père, soupe	p (occlusive)
b	bon, robe	b (occlusive)
t	terre, vite	t (occlusive)
d	dans, aide	d (occlusive)
k	carré, laque	k (occlusive)
g	gare, bague	g (occlusive)
f	feu, neuf	f (fricative)
v	vous, rêve	v (fricative)
s	sale, dessous	s (fricative)
z	zéro, maison	z (fricative)
S	chat, tâche	ch (fricative)
Z	gilet, mijoter	ge (fricative)
m	main, femme	m (cons. nasale)
n	nous, tonne	n (cons. nasale)
N	agneau, vigne	gn (c. nasale palat.)
l	lent, sol	l (liquide)
R	rue, venir	R
x	jota	jota (emprunt espagn.)
G	camping	ng (emprunt angl.)

Cat Gram (cat Gram)
La catégorie grammaticale donnée par Lexique. Une même entrée orthographique ayant deux catégorie grammaticales distinctes peut avoir deux représentations phonologiques différentes et donc deux associations graphème-phonème différentes (p.ex. ils ferment vs. le ferment).

Décomp Graphemes (grapheme)
Segmentation (séparateur = ".") d'une entrée orthographique en ses différents graphèmes (ex : château -> ch.â.t.eau)

Décomp Association Graphème-Phonème (assoc)
Segmentation d'une entrée orthographique en ses différentes associations graphème-phonème (ex : ch-S.â-a.t-t.eau-o). L'association entre un graphème et un phonème est représentée par "-". Les différentes associations sont séparées par ".".

Consistance Graphèmes-Phonèmes | Décomposition | Type (decomp)
Pour un mot donné la décomposition de toutes les consistances graphèmes-phonèmes (château -> 0.91 1.00 0.93 1.00). Dans ce cas particulier ch-S a une consistance de .91 en position initiale, â-a a une consistance de 1 en position milieu, t-t a une consistance de .93 en position milieu, eau-o a une consistance de 1 en position finale.
Par exemple, supposons que dans la position initiale, le graphème "ch" apparaisse 15 fois dans notre corpus, le phonème /S/ apparaît 20 fois, la paire ch-/S/ apparaît 10 fois dans notre corpus et la paire ch-/k/ apparaît 5 fois. Pour la paire "ch-/S/" sa **fréquence d'association graphème-phonème** sera de 10/15 = 67%, alors que sa fréquence d'association phonème-graphème sera de 10/20 = 50%. Pour la **fréquence de type** chaque entrée orthographique ne sera compté qu'une fois (alors que pour la **fréquence de token** la fréquence de chaque entrée orthographique sera prise en compte).

Consistance Graphèmes-Phonèmes | Décomposition | Token (decomp)
voir ci-dessus mais calculé par token. Ainsi chaque calcul de consistance a été pondéré par la fréquence du mot.

Consistance Graphèmes-Phonèmes | Moyenne | Type (Freq_GP)
Moyenne des consistances graphèmes-phonèmes par type constituant le mot (château -> (0.91 + 1 + 0.93 +1) / 4 = 0.96

Consistance Graphèmes-Phonèmes | Moyenne | Token (FreqTok_GP)
Idem par token

Consistance Graphèmes-Phonèmes | Initiale | Type (Freq_I_GP)
La consistance graphème-phonème pour le graphème **initial** du mot. (château -> ch-S en position initiale -> .91)

Consistance Graphèmes-Phonèmes | Initiale | Token (Freq_I_Tok_GP)

Idem par token

Consistance Graphèmes-Phonèmes | Milieu | Type (Freq_M_GP)
La moyenne des consistances graphème-phonème pour les graphème **en milieu** du mot. (château -> â-/a/ + t-/t/ [les 2 en milieu de mot] -> (1 + 0.93) / 2 = 0.96)

Consistance Graphèmes-Phonèmes | Milieu | Token (Freq_M_Tok_GP)
Idem par token

Consistance Graphèmes-Phonèmes | Finale | Type (Freq_F_GP)
La consistance graphème-phonème pour le graphème **final** du mot. (château -> eau-o en **position finale** -> .91)

Consistance Graphèmes-Phonèmes | Finale | Token (Freq_F_Tok_GP)
Idem par token

Consistance Graphèmes-Phonèmes | Freq Mini | Type (minfreqgraph_GP)
Consistance Graphèmes-Phonèmes la plus faible dans le mot (château -> 0.91 1.00 0.93 1.00 -> .91)

Consistance Graphèmes-Phonèmes | Freq Mini | Token (minfreqgraphTok_GP)
Idem par token

Régularité Graphèmes-Phonèmes | Décomposition | Type (regTy-GP)
Décomposition de l'entrée orthographique en association graphème-phonème régulière (1) ou irrégulière (0). Une association graphème-phonème est considérée comme régulière si cette association est la plus fréquente.
Par exemple, le mot "femme" donne "1.0.1.1." : toutes les associations graphème-phonème sont régulières sauf l'association "e-a".

Régularité Graphèmes-Phonèmes | Décomposition | Token (regTo-GP)
Idem par token

Régularité Graphèmes-Phonèmes | Nb irrégularités | Type (countregTy_GP)
Nombre d'irrégularités de l'entrée orthographique **(**château -> 1.1.1.1. -> 4)

Régularité Graphèmes-Phonèmes | Nb irrégularités | Token (countregTo_GP)
Idem par token

Régularité Graphèmes-Phonèmes | Pos 1ère irrégularité | Type (posregTy_GP)
Position dans le mot de la 1ère irrégularité constatée. (femme -> f-f.e-a.mm-m.e-# -> 1.0.1.1. -> 2)

Régularité Graphèmes-Phonèmes | Pos 1ère irrégularité | Token (posregTo_GP)
Idem par token

Consistance Phonèmes-Graphèmes | Décomposition | Type (decomp)
Pour un mot donné la décomposition de toutes les consistances phonèmes-graphèmes (château -> .92 .03 .91 .22). Dans ce cas particulier ch-S a une consistance de .92 en position initiale, â-a a une consistance de .03 en position milieu (car le phonème /a/ s'écrit rarement "â"), t-t a une consistance de .91 en position milieu, eau-o a une consistance de .22 en position finale.
Par exemple, supposons que dans la position initiale, le graphème "ch" apparaisse 15 fois dans notre corpus, le phonème /S/ apparaît 20 fois, la paire ch-/S/ apparaît 10 fois dans notre corpus et la paire ch-/k/ apparaît 5 fois. Pour la paire "ch-/S/" sa fréquence d'association graphème-phonème sera de 10/15 = 67%, alors que sa **fréquence d'association phonème-graphème** sera de 10/20 = 50%. Pour la **fréquence de type** chaque entrée orthographique ne sera compté qu'une fois (alors que pour la **fréquence de token** la fréquence de chaque entrée orthographique sera prise en compte).

Consistance Phonèmes-Graphèmes | Décomposition | Token (decomp)
Idem par token

Consistance Phonèmes-Graphèmes | Moyenne | Type (Freq_GP)
Moyenne des consistances phonèmes-graphèmes par type constituant le mot (château -> (0.92 + 0.03 + 0.91 + 0.22) / 4 = 0.52

Consistance Phonèmes-Graphèmes | Moyenne | Token (FreqTok_GP)
Idem par token

Consistance Phonèmes-Graphèmes | Initiale | Type (Freq_I_GP)
La consistance phonèmes-graphèmes pour le phonème **initial** du mot. (château -> ch-S en position initiale -> .92)

Consistance Phonèmes-Graphèmes | Initiale | Token (Freq_I_Tok_GP)
Idem par token

Consistance Phonèmes-Graphèmes | Milieu | Type (Freq_M_GP)
La moyenne des consistances graphème-phonème pour les phonèmes **en milieu** du mot. (château -> â-/a/ + t-/t/ [les 2 en milieu de mot] -> (.03 + 0.91) / 2 = .47)

Consistance Phonèmes-Graphèmes | Milieu | Token (Freq_M_Tok_GP)
Idem par token

Consistance Phonèmes-Graphèmes | Finale | Type (Freq_F_GP)

La consistance phonème-graphème pour le phonème **final** du mot. (château -> eau-o en **position finale** -> .22)

Consistance Phonèmes-Graphèmes | Finale | Token (Freq_F_Tok_GP)
Idem par token

Consistance Phonèmes-Graphèmes | Freq Mini | Type (minfreqgraph_GP)
Consistance Phonèmes-Graphèmes la plus faible dans le mot (château -> 0.91 1.00 0.93 1.00 -> .91)

Consistance Phonèmes-Graphèmes | Freq Mini | Token (minfreqgraphTok_GP)
Idem par token

Régularité Phonèmes-Graphèmes | Décomposition | Type (regTy-GP)
Décomposition de l'entrée phonologique en association phonème-graphème régulière (1) ou irrégulière (0). Une association phonème-graphème est considérée comme régulière si cette association est la plus fréquente.
Par exemple, le mot "château" donne "1.0.1.0." : toutes les associations graphème-phonème sont régulières sauf l'association "e-a".

Régularité Phonèmes-Graphèmes | Décomposition | Token (regTo-GP)
Idem par token

Régularité Phonèmes-Graphèmes | Nb irrégularités | Type (countregTy_GP)
Nombre d'irrégularités de l'entrée orthographique **(**château ->1.0.1.0. -> 2)

Régularité Phonèmes-Graphèmes | Nb irrégularités | Token (countregTo_GP)
Idem par token

Régularité Phonèmes-Graphèmes | Pos 1ère irrégularité | Type (posregTy_GP)
Position dans le mot de la 1ère irrégularité constatée. (château ->1.0.1.0. -> 2)

Régularité Phonèmes-Graphèmes | Pos 1ère irrégularité | Token (posregTo_GP)
Idem par token

Complexité moyenne des graphèmes (complexmoygraph)
Nombre de lettres moyen des graphèmes composant l'entrée orthographique (ch.â.t.eau -> (2+1+1+3)/4)

Fichier Lexique.Infra.Freq.Let.Bigr.Trig.Syl.Phon.Biph.xlsb

Item (item)
Les formes orthographiques utilisées extraites de Lexique.

Phonologie (phono)
La forme phonologique donnée par Lexique (voir fichier précédent).

Categ Gram (cgram)
La catégorie grammaticale donnée par Lexique. Une même entrée orthographique ayant deux catégorie grammaticales distinctes peut avoir deux représentations phonologiques différentes et donc deux associations graphème-phonème différentes (p.ex. ils ferment vs. le ferment).

Décomposition Graphèmes (Graphèmes)
Segmentation (séparateur = ".") d'une entrée orthographique en ses différents graphèmes (ex : château -> ch.â.t.eau)

SylPhono (SylPhono)
Décomposition de la représentation phonologique en syllabes (château -> /Sato/ -> /Sa-to/).

Fréquence Lettres | Décomp | Type (LetDecompTy)
Fréquence de chaque lettre constituant le mot (château -> c.h.â.t.e.a.u. 15186-12780-2157-57330-98768-87269-842)

Fréquence Lettres | Freq Moy | Type (LetFreqTy)
Fréquence moyenne des lettres constituant le mot (château -> c.h.â.t.e.a.u. 15186-12780-2157-57330-98768-87269-842 -> 15186+12780+2157+57330+98768+87269+842 / 7 = 39190

Fréquence Lettres | Décomp | Token (LetDecompTo)
Idem par token

Fréquence Lettres | Freq Moy | Token (LetFreqTo)
Idem par token

Les calculs correspondant aux 4 champs précédents ont été également réalisés pour les unités infra-lexicales suivantes :
bigrammes : séquences de deux lettres successives dans un mot (château a 6 bigrammes -> ch.hâ.ât.te.ea.au)
trigrammes : séquences successives de trois lettres dans un mot (château a 5 trigrammes -> châ.hât.âte.tea.eau
phonèmes : plus petites unités sonores permettant de distinguer 2 mots (château a 4 phonèmes -> /S.a.t.o/)
biphones : séquences de deux phonèmes successifs dans un mot (château a 3 biphones -> /Sa/./at/./to/)
graphèmes : représentations écrites des phonèmes (château a 4 graphèmes -> ch.â.t.eau)
syllabes phonologiques : représentations phonologiques des syllabes (château -> /Sa/./to/)