

SyllabO+ :

Description des fichiers téléchargeables (valeurs et formules de calcul)

Version du 4 aout 2016



LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY



UNIVERSITÉ
LAVAL

Valeurs figurant dans les tableaux, avec définitions et détails des calculs effectués

Les fichiers téléchargeables sont des fichiers tableurs, contenant des données sur chaque ligne séparées par un caractère de séparation (une tabulation). Ces fichiers contiennent de nombreuses colonnes, lesquelles sont décrites ici. Les syllabes et phones sont notés en alphabet phonétique international (API). Pour des conseils pour ouvrir ces fichiers et gérer les caractères spéciaux qu'ils contiennent, référez-vous au guide d'utilisation.

Exemple de fichier :

Syllabe	Structure	Fréquence	Pourcentage	Rang centile
a	V	8994	2.962099356	99.98081719
se	CV	6497	2.139733101	99.96163438
de	CV	5245	1.727397278	99.94245156
də	CV	5156	1.698085866	99.92326875
le	CV	4802	1.5814989	99.90408594
e	V	4510	1.48533112	99.88490313
la	CV	4133	1.361169295	99.86572031
la	CV	4107	1.35260641	99.8465375
mã	CV	3763	1.239312861	99.82735469
kə	CV	3683	1.212965525	99.80817188
ã	V	3598	1.184971479	99.78898907

• *Syllabe / Paire / Triade / Tétrade*

Transcription de la syllabe, paire, triade ou tétrade en alphabet phonétique international

• *Structure*

Composition de la syllabe selon les consonnes et les voyelles

(C = consonnes, V = voyelles, S = semi-voyelles)

Consonnes : [p] [t] [k] [b] [d] [g] [f] [s] [ʃ] [v] [z] [ʒ] [m] [n] [ɲ] [ŋ] [l] [r] [ɹ] [ɻ] [ð] [θ] [h]* [x]**

Voyelles : [i] [y] [u] [e] [ø] [o] [ə] [ɛ] [œ] [ɔ] [a] [ɑ] [ɛ̃] [ã] [õ] [œ̃] [ʌ] [ɒ] [ɜ] [æ] [ɪ] [ʏ] [ʊ]*

Semi-voyelles : [w] [j] [ɥ]

À noter que le symbole # correspond aux sons inintelligibles

* Utilisés seulement lorsque le locuteur emploie une prononciation anglaise.

** Utilisés seulement lorsque le locuteur emploie une prononciation espagnole.

- **Fréquence**

Nombre total d'occurrences de la syllabe, paire, triade ou tétrade (valeur absolue) dans le corpus.

- **Pourcentage**

Fréquence de la syllabe, paire, triade ou tétrade dans le corpus, en pourcentage

Calcul : (fréquence / nombre total d'unités [syllabes / paires / triades / tétrades]) * 100

- **Rang centile**

Rang centile de la syllabe, paire, triade ou tétrade dans le corpus

Calcul : effectué à l'aide de la fonction *percentileofscore* (*kind* = 'strict') de la librairie *scipy*

(*stats*) dans un script *Python* - Voir explications ci-dessous

Le rang centile est une mesure de position de valeur utilisée en statistiques. Il indique le pourcentage des données dont la valeur est inférieure à la donnée observée.

Pour plus d'informations sur le calcul effectué à l'aide de la fonction *percentileofscore* de la librairie *scipy*, veuillez consulter la documentation suivante.

<http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.percentileofscore.html>

- **Probabilité de transition avant**

Probabilité que la première syllabe d'une paire soit suivie de la deuxième syllabe

Calcul : (fréquence de la paire / fréquence de la première syllabe) * 100

- **Probabilité de transition arrière**

Probabilité que la deuxième syllabe d'une paire soit précédée de la première syllabe

Calcul : (fréquence de la paire / fréquence de la deuxième syllabe) * 100

- **Score d'association « pointwise mutual information » (PMI)**

Mesure d'association entre les éléments d'une paire ou d'une triade

Calcul : effectué à l'aide de la fonction *pmi* de la librairie *nltk* (*collocations* –

BigramsAssocMeasures ou *TrigramsAssocMeasures*) dans un script *Python* - Voir explications ci-dessous (section suivante)

- **Variante du score d'association « mutual information » (MI-like)**

Variante de la mesure d'association entre les éléments d'une paire ou d'une triade

Calcul : effectué à l'aide de la fonction *mi_like* de la librairie *nltk* (*collocations* –

BigramsAssocMeasures ou *TrigramsAssocMeasures*) dans un script *Python* - Voir explications ci-dessous

Les scores d'association – que ce soit le *pointwise mutual information (PMI)*, le *mutual information (MI)* ou une variante de celui-ci – sont des mesures permettant de déterminer statistiquement la dépendance mutuelle entre des valeurs.

Le score d'association *pointwise mutual information* permet de mesurer l'information commune (association) entre deux valeurs particulières des distributions.

$$pmi(x; y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Le *MI-like* est une variante du score d'association *mutual information*. Il permet également de mesurer l'information commune (association) entre deux valeurs, tout en accordant une moins grande importance aux événements rares (contrairement au *PMI*, qui calculera un score élevé pour les éléments rares). Le *MI-like* correspond au *mutual information* avec le numérateur au cube.

$$mi_like(x; y) = \frac{(p(x, y))^3}{p(x)p(y)}$$

Voici deux exemples pour illustrer concrètement une différence entre un score *PMI* et un score *MI-like*. La paire fréquente [vu za] (0,055%) est associée à un score *PMI* de **5,81** et un score *MI-like* de **4,95** dans notre corpus. En contraste, la paire peu fréquente [kam pys] (0,001%) obtient un score très élevé de *PMI* de **12,92** et un score beaucoup plus faible de *MI-like* de **0,23**, reflétant que la paire, bien que fortement associée, est extrêmement rare.

Pour plus d'informations sur le calcul effectué à l'aide de la fonction *pmi* ou *mi_like* de la librairie *nltk* (*collocations – BigramsAssocMeasures* ou *TrigramsAssocMeasures*), consultez la documentation suivante, aux entrées « *def pmi* » et « *def mi_like* ».

http://www.nltk.org/_modules/nltk/metrics/association.html

Tableau des syllabes

- *Syllabe*
- *Structure*
- *Fréquence*
- *Pourcentage*
- *Rang centile*

Tableau des paires de syllabes

- **Paire**
- **Fréquence** (paire)
- **Pourcentage** (paire)
- **Rang centile** (paire)
- **Probabilité de transition avant** (paire)
- **Probabilité de transition arrière** (paire)
- **Score d'association « pointwise mutual information »** (paire)
- **Variante du score d'association « mutual information »** (paire)
- **1^{ère} syllabe**
- **Structure** (1^{ère} syllabe)
- **Fréquence** (1^{ère} syllabe)
- **Pourcentage** (1^{ère} syllabe)
- **Rang centile** (1^{ère} syllabe)
- **2^e syllabe**
- **Structure** (2^e syllabe)
- **Fréquence** (2^e syllabe)
- **Pourcentage** (2^e syllabe)
- **Rang centile** (2^e syllabe)

Tableau des triades

- **Triade**
- **Fréquence** (triade)
- **Pourcentage** (triade)
- **Rang centile** (triade)
- **Score d'association « pointwise mutual information »** (triade)
- **Variante du score d'association « mutual information »** (triade)
- **Probabilité de transition avant** (paire syllabes 1 – 2)
- **Probabilité de transition arrière** (paire syllabes 1 – 2)
- **Score d'association « pointwise mutual information »** (paire syllabes 1 – 2)
- **Variante du score d'association « mutual information »** (paire syllabes 1 – 2)
- **Probabilité de transition avant** (paire syllabes 2 – 3)
- **Probabilité de transition arrière** (paire syllabes 2 – 3)
- **Score d'association « pointwise mutual information »** (paire syllabes 2 – 3)
- **Variante du score d'association « mutual information »** (paire syllabes 2 – 3)
- **1^{ère} syllabe**
- **Structure** (1^{ère} syllabe)
- **Fréquence** (1^{ère} syllabe)
- **Pourcentage** (1^{ère} syllabe)
- **Rang centile** (1^{ère} syllabe)
- **2^e syllabe**
- **Structure** (2^e syllabe)
- **Fréquence** (2^e syllabe)
- **Pourcentage** (2^e syllabe)

- **Rang centile** (2^e syllabe)
- **3^e syllabe**
- **Structure** (3^e syllabe)
- **Fréquence** (3^e syllabe)
- **Pourcentage** (3^e syllabe)
- **Rang centile** (3^e syllabe)

Tableau des tétrades

- **Tétrade**
- **Fréquence** (tétrade)
- **Pourcentage** (tétrade)
- **Rang centile** (tétrade)
- **Probabilité de transition avant** (paire syllabes 1 – 2)
- **Probabilité de transition arrière** (paire syllabes 1 – 2)
- **Score d'association « pointwise mutual information »** (paire syllabes 1 – 2)
- **Variante du score d'association « mutual information »** (paire syllabes 1 – 2)
- **Probabilité de transition avant** (paire syllabes 2 – 3)
- **Probabilité de transition arrière** (paire syllabes 2 – 3)
- **Score d'association « pointwise mutual information »** (paire syllabes 2 – 3)
- **Variante du score d'association « mutual information »** (paire syllabes 2 – 3)
- **Probabilité de transition avant** (paire syllabes 3 – 4)
- **Probabilité de transition arrière** (paire syllabes 3 – 4)
- **Score d'association « pointwise mutual information »** (paire syllabes 3 – 4)
- **Variante du score d'association « mutual information »** (paire syllabes 3 – 4)
- **1^{ère} syllabe**
- **Structure** (1^{ère} syllabe)
- **Fréquence** (1^{ère} syllabe)
- **Pourcentage** (1^{ère} syllabe)
- **Rang centile** (1^{ère} syllabe)
- **2^e syllabe**
- **Structure** (2^e syllabe)

- **Fréquence** (2^e syllabe)
- **Pourcentage** (2^e syllabe)
- **Rang centile** (2^e syllabe)
- **3^e syllabe**
- **Structure** (3^e syllabe)
- **Fréquence** (3^e syllabe)
- **Pourcentage** (3^e syllabe)
- **Rang centile** (3^e syllabe)
- **4^e syllabe**
- **Structure** (4^e syllabe)
- **Fréquence** (4^e syllabe)
- **Pourcentage** (4^e syllabe)
- **Rang centile** (4^e syllabe)