SyllabO+ :

Description des fichiers téléchargeables (valeurs et formules de calcul)

Version 20 novembre 2017

LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY

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 les 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
lɑ	CV	4133	1.361169295	99.86572031
la	CV	4107	1.35260641	99.8465375
mɑ̃	CV	3763	1.239312861	99.82735469
kə	CV	3683	1.212965525	99.80817188
ɑ̃	V	3598	1.184971479	99.78898907

• ***Phone / Syllabe / Mot / Lemme / Paire / Triade***

Transcription de l'unité (phone, syllabe, mot ou lemme) ou de la séquence d'unités (paire ou triade) en alphabet phonétique international (lorsque applicable)

• ***Structure***

Composition de la syllabe ou du phone 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] [ð] [θ] [h]* [x]**

Voyelles : [i] [y] [u] [e] [ø] [o] [ə] [ɛ] [œ] [ɔ] [a] [ɑ] [ɛ̃] [ɑ̃] [ɔ̃] [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 l'unité (phone, syllabe, mot ou lemme) ou de la séquence d'unités (paire ou triade) en valeur absolue dans le corpus.

• ***Pourcentage***

Fréquence de l'unité (phone, syllabe, mot ou lemme) ou de la séquence d'unités (paire ou triade) dans le corpus, en pourcentage

Calcul : (fréquence / nombre total d'unités) * 100

• ***Rang centile***

Rang centile de l'unité (phone, syllabe, mot ou lemme) ou de la séquence d'unités (paire ou triade) 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 unité (phone, syllabe, mot ou lemme) d'une paire soit suivie de la deuxième unité

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

• ***Probabilité de transition arrière***

Probabilité que la deuxième unité (phone, syllabe, mot ou lemme) d'une paire soit précédée de la première unité

Calcul : (fréquence de la paire / fréquence de la deuxième unité) * 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.

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

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 simialire de *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**, refétant 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

Voir pages suivantes pour la liste des colonnes dans chaque tableau, soit les :

• Phones

• Paires de phones

• Syllabes

• Paires de syllabes

• Triades de syllabes

• Mots/lemmes

• Paires de mots/lemmes

• Triades de mots/lemmes

Tableau des phones

- ***Phone***
- ***Structure***
- ***Fréquence***
- ***Pourcentage***
- ***Rang centile***

Tableau des diphones

- ***Paire de phones (diphone)***
- ***Structure totale***
- ***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)*
- ***1er phone***
- ***Structure*** *(1er phone)*
- ***Fréquence*** *(1er phone)*
- ***Pourcentage*** *(1er phone)*
- ***Rang centile*** *(1er phone)*
- ***2^{e} phone***
- ***Structure*** *(2^{e} phone)*
- ***Fréquence*** *(2^{e} phone)*
- ***Pourcentage*** *(2^{e} phone)*
- ***Rang centile*** *(2^{e} phone)*

Tableau des syllabes

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

Tableau des paires de syllabes

- ***Paire de syllabes***
- ***Structure totale***
- ***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)*
- ***1ère syllabe***
- ***Structure*** *(1ère syllabe)*
- ***Fréquence*** *(1ère syllabe)*
- ***Pourcentage*** *(1ère syllabe)*
- ***Rang centile*** *(1ère syllabe)*
- ***2e syllabe***
- ***Structure*** *(2e syllabe)*
- ***Fréquence*** *(2e syllabe)*
- ***Pourcentage*** *(2e syllabe)*
- ***Rang centile*** *(2e syllabe)*

Tableau des triades de syllabes

- ***Triade de syllabes***
- ***Structure totale***
- ***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)*
- ***1ère syllabe***
- ***Structure*** *(1ère syllabe)*
- ***Fréquence*** *(1ère syllabe)*
- ***Pourcentage*** *(1ère syllabe)*
- ***Rang centile*** *(1ère syllabe)*
- ***2e syllabe***
- ***Structure*** *(2e syllabe)*
- ***Fréquence*** *(2e syllabe)*

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

Tableau des mots/lemmes

- ***Mot/lemme***
- ***Fréquence***
- ***Pourcentage***
- ***Rang centile***

Tableau des paires de mots/lemmes

- ***Paire de mots/lemmes***
- ***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)*
- ***1er mot/lemme***
- ***Fréquence*** *(1er mot/lemme)*
- ***Pourcentage*** *(1er mot/lemme)*
- ***Rang centile*** *(1er mot/lemme)*
- ***2e mot/lemme***
- ***Fréquence*** *(2e mot/lemme)*
- ***Pourcentage*** *(2e mot/lemme)*
- ***Rang centile*** *(2e mot/lemme)*

Tableau des triades de mots/lemmes

- ***Triade de syllabes***
- ***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 mots/lemmes 1 – 2)*
- ***Probabilité de transition arrière*** *(paire mots/lemmes 1 – 2)*
- ***Score d'association « pointwise mutual information »*** *(paire mots/lemmes 1 – 2)*
- ***Variante du score d'association « mutual information »*** *(paire mots/lemmes 1 – 2)*
- ***Probabilité de transition avant*** *(paire mots/lemmes 2 – 3)*
- ***Probabilité de transition arrière*** *(paire mots/lemmes 2 – 3)*
- ***Score d'association « pointwise mutual information »*** *(paire mots/lemmes 2 – 3)*
- ***Variante du score d'association « mutual information »*** *(paire mots/lemmes 2 – 3)*
- ***1er mot/lemme***
- ***Fréquence*** *(1er mot/lemme)*
- ***Pourcentage*** *(1er mot/lemme)*
- ***Rang centile*** *(1er mot/lemme)*
- ***2^{e} mot/lemme***
- ***Fréquence*** *(2^{e} mot/lemme)*
- ***Pourcentage*** *(2^{e} mot/lemme)*
- ***Rang centile*** *(2^{e} mot/lemme)*
- ***3^{e} mot/lemme***
- ***Fréquence*** *(3^{e} mot/lemme)*

• ***Pourcentage*** *(3e mot/lemme)*

• ***Rang centile*** *(3e mot/lemme)*