SyllabO+:
Output file description (values and formulas)

Version August 4th 2016

LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY

UNIVERSITÉ
LAVAL

Description of the output files

The output files are tab-delineated files with a number of columns that are described here.

Example of a file:

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

• ***Syllable / Pair / Triad / Tetrad***

Transcription of the syllable, pair (group of 2 syllables), triad (group of 3 syllables) or tetrad (group of 4 syllables) in International Phonetic Alphabet

• ***Syllable structure***

Composition of the syllable, according to consonants and vowels
(C = consonants, V = vowels, S = semi-vowels)

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

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

Semi-vowels: [w] [j] [ɥ]

Note that symbol # corresponds to unintelligible sounds

* Used only when speaker uses an English pronunciation.

** Used only when speaker uses a Spanish pronunciation.

• ***Frequency***

Total number of occurrences (absolute value) of the syllable, pair, triad or tetrad in the corpus

• ***Percentage***

Frequency of the syllable, pair, triad or tetrad in the corpus, in percentage

Calculation: (frequency / total number of units [syllable / pair / triad / tetrad]) * 100

• ***Percentile of score***

Percentile of the syllable, pair, triad or tetrad in the corpus

Calculation: executed by the *percentileofscore (kind = 'strict')* function of the *scipy* library *(stats)* in a *Python* script - *See explanations below*

Percentile of score is a measure of position used in statistics. It indicates the percentage of data whose value is lower than the observed data.

For more information on the calculation performed by the *percentileofscore* function of the *scipy* library, see the following documentation.

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

• ***Forward transition probability***

Probability that the first syllable of a pair would be followed by the second syllable

Calculation: (frequency of the pair / frequency of the first syllable) * 100

• ***Backward transition probability***

Probability that the second syllable of a pair would be preceded by the first syllable

Calculation: (frequency of the pair / frequency of the second syllable) * 100

• ***Pointwise mutual information (PMI)***

Association measure between elements of a pair or a triad

Calculation: executed by the *pmi* function of the *nltk* library *(collocations – BigramsAssocMeasures or TrigramsAssocMeasures)* in a *Python* script - *See explanations below (next section)*

• ***Variant of mutual information (MI-like)***

Variant of the association measure between elements of a pair or a triad

Calculation: executed by the *mi_like* function of the *nltk* library *(collocations – BigramsAssocMeasures or TrigramsAssocMeasures)* in a *Python* script - *See explanations below*

Association scores – whether *pointwise mutual information (PMI)*, *mutual information (MI)* or its variants – are measures that determine the mutual dependency between values.

The PMI enables the calculation of common information (association) between two particular values of a distribution.

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

MI-like is a variant of MI. It also enables the calculation of common information (association) between two values, but it gives less importance to rare events (unlike *PMI*, which calculates a high score for rare events). *MI-like* corresponds to *MI* with the numerator cubed.

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

Here is an illustration of the difference between *PMI* and *MI-like* scores. The frequent pair [vu za] (0,055%) has a *PMI* score of **5.81** in our database and a similar *MI-like* score of **4.95**. In contrast, the infrequent pair [kam pys] (0,001%) obtains a *PMI* score of **12.92** and a much lower *MI-like* score of only **0.23**, reflecting the frequency of the pair. This shows that the frequency of the pair itself has an impact on the calculation of *MI-like* but not *PMI*.

For more information on the calculation performed by the *pmi* or *mi_like* function of the *nltk* library *(collocations – BigramsAssocMeasures or TrigramsAssocMeasures)*, see the following documentation, at entries “def pmi” and “def mi_like”.

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

Syllables table

- ***Syllable***
- ***Structure***
- ***Frequency***
- ***Percentage***
- ***Percentile of score***

Pairs table

- ***Pair***
- ***Frequency*** *(pair)*
- ***Percentage*** *(pair)*
- ***Percentile of score*** *(pair)*
- ***Forward transition probability*** *(pair)*
- ***Backward transition probability*** *(pair)*
- ***Pointwise mutual information*** *(pair)*
- ***Variant of mutual information*** *(pair)*
- ***1st syllable***
- ***Structure*** *(1st syllable)*
- ***Frequency*** *(1st syllable)*
- ***Percentage*** *(1st syllable)*
- ***Percentile of score*** *(1st syllable)*
- ***2nd syllable***
- ***Structure*** *(2nd syllable)*
- ***Frequency*** *(2nd syllable)*
- ***Percentage*** *(2nd syllable)*
- ***Percentile of score*** *(2nd syllable)*

Triads table

- ***Triad***
- ***Frequency*** *(triad)*
- ***Percentage*** *(triad)*
- ***Percentile of score*** *(triad)*
- ***Pointwise mutual information*** *(triad)*
- ***Variant of mutual information*** *(triad)*
- ***Forward transition probability*** *(pair syllables 1 – 2)*
- ***Backward transition probability*** *(pair syllables 1 – 2)*
- ***Pointwise mutual information*** *(pair syllables 1 – 2)*
- ***Variant of mutual information*** *(pair syllables 1 – 2)*
- ***Forward transition probability*** *(pair syllables 2 – 3)*
- ***Backward transition probability*** *(pair syllables 2 – 3)*
- ***Pointwise mutual information*** *(pair syllables 2 – 3)*
- ***Variant of mutual information*** *(pair syllables 2 – 3)*
- ***1st syllable***
- ***Structure*** *(1st syllable)*
- ***Frequency*** *(1st syllable)*
- ***Percentage*** *(1st syllable)*
- ***Percentile of score*** *(1st syllable)*
- ***2nd syllable***
- ***Structure*** *(2nd syllable)*
- ***Frequency*** *(2nd syllable)*
- ***Percentage*** *(2nd syllable)*
- ***Percentile of score*** *(2nd syllable)*
- ***3rd syllable***
- ***Structure*** *(3rd syllable)*
- ***Frequency*** *(3rd syllable)*
- ***Percentage*** *(3rd syllable)*
- ***Percentile of score*** *(3rd syllable)*

Tetrads table

- ***Tetrad***
- ***Frequency*** *(tetrad)*
- ***Percentage*** *(tetrad)*
- ***Percentile of score*** *(tetrad)*
- ***Forward transition probability*** *(pair syllables 1 – 2)*
- ***Backward transition probability*** *(pair syllables 1 – 2)*
- ***Pointwise mutual information*** *(pair syllables 1 – 2)*
- ***Variant of mutual information*** *(pair syllables 1 – 2)*
- ***Forward transition probability*** *(pair syllables 2 – 3)*
- ***Backward transition probability*** *(pair syllables 2 – 3)*
- ***Pointwise mutual information*** *(pair syllables 2 – 3)*
- ***Variant of mutual information*** *(pair syllables 2 – 3)*
- ***Forward transition probability*** *(pair syllables 3 – 4)*
- ***Backward transition probability*** *(pair syllables 3 – 4)*
- ***Pointwise mutual information*** *(pair syllables 3 – 4)*
- ***Variant of mutual information*** *(pair syllables 3 – 4)*
- ***1st syllable***
- ***Structure*** *(1st syllable)*
- ***Frequency*** *(1st syllable)*
- ***Percentage*** *(1st syllable)*
- ***Percentile of score*** *(1st syllable)*
- ***2nd syllable***
- ***Structure*** *(2nd syllable)*
- ***Frequency*** *(2nd syllable)*
- ***Percentage*** *(2nd syllable)*
- ***Percentile of score*** *(2nd syllable)*
- ***3rd syllable***
- ***Structure*** *(3rd syllable)*
- ***Frequency*** *(3rd syllable)*
- ***Percentage*** *(3rd syllable)*
- ***Percentile of score*** *(3rd syllable)*

- ***4th syllable***
- ***Structure*** *(4th syllable)*
- ***Frequency*** *(4th syllable)*
- ***Percentage*** *(4th syllable)*
- ***Percentile of score*** *(4th syllable)*