SyllabO+:
Output file description (values and formulas)

Version ~~du 20~~ November 20, 2017

LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY

Description of the output files

The output files are tab-delineated files with a number of columns that are described here. Syllables and phones are transcribed in international phonetic alphabet (IPA). For advice on opening the files and managing special characters, please refer to user manual.

Example of a file:

Syllable	Structure	Frequency	Percentage	Percentile rank
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 / Syllable / Word / Lemma / Pair / Triad***

Transcription of the unit (phone, syllable, word, or lemma) or sequence of units (pair or triad) in International Phonetic Alphabet (when applicable)

• ***Structure***

Composition of the syllable or phone, 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 unit (phone, syllable, word, lemma) or sequence of units (pair, triad) in the corpus

• ***Percentage***

Frequency of the unit (phone, syllable, word, lemma) or sequence of units (pair, triad) in the corpus, in percentage

Calculation: (frequency / total number of units) * 100

• ***Percentile rank***

Percentile rank of the unit (phone, syllable, word, lemma) or sequence of units (pair, triad) in the corpus

Calculation: executed by the *percentiledescore (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 *percentiledescore* function of the *scipy* library, see the following documentation.

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

• ***Forward transition probability***

Probability that the first unit (phone, syllable, word, lemma) of a pair would be followed by the second unit

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

• ***Backward transition probability***

Probability that the second unit (phone, syllable, word, lemma) of a pair would be preceded by the first unit

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.

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

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

See the following pages for the list of all the columns present in each ~~different~~ table:

- Phones
- Pairs of phones
- Syllables
- Pairs of syllables
- Triads of syllables
- Words/lemmas
- Pairs of words/lemmas
- Triads of words/lemmas

Phones table

- ***Phone***
- ***Structure***
- ***Frequency***
- ***Percentage***
- ***Percentile rank***

Pairs of phones table

- ***Pair of phones (diphone)***
- ***Total structure***
- ***Frequency*** *(pair)*
- ***Percentage*** *(pair)*
- ***Percentile rank*** *(pair)*
- ***Transition probability forward*** *(pair)*
- ***Transition probability backward*** *(pair)*
- ***Pointwise mutual information (PMI)*** *(pair)*
- ***Variant of mutual information (MI-like)*** *(pair)*
- ***1st phone***
- ***Structure*** *(1st phone)*
- ***Frequency*** *(1st phone)*
- ***Percentage*** *(1st phone)*
- ***Percentile rank*** *(1st phone)*
- ***2nd phone***
- ***Structure*** *(2nd phone)*
- ***Frequency*** *(2nd phone)*
- ***Percentage*** *(2nd phone)*
- ***Percentile rank*** *(2nd phone)*

Syllables table

- ***Syllable***
- ***Structure***
- ***Frequency***
- ***Percentage***
- ***Percentile rank***

Pairs of syllables table

- ***Pair of syllables***
- ***Total structure***
- ***Frequency*** *(pair)*
- ***Percentage*** *(pair)*
- ***Percentile rank*** *(pair)*
- ***Transition probability forward*** *(pair)*
- ***Transition probability backward*** *(pair)*
- ***Pointwise mutual information (PMI)*** *(pair)*
- ***Variant of mutual information (MI-like)*** *(pair)*
- ***1st syllable***
- ***Structure*** *(1st syllable)*
- ***Frequency*** *(1st syllable)*
- ***Percentage*** *(1st syllable)*
- ***Percentile rank*** *(1st syllable)*
- ***2nd syllable***
- ***Structure*** *(2nd syllable)*
- ***Frequency*** *(2nd syllable)*
- ***Percentage*** *(2nd syllable)*
- ***Percentile rank*** *(2nd syllable)*

Triads of syllables table

- ***Triad of syllables***
- ***Total structure***
- ***Frequency*** *(triad)*
- ***Percentage*** *(triad)*
- ***Percentile rank*** *(triad)*
- ***Pointwise mutual information (PMI)*** *(triad)*
- ***Variant of mutual information (MI-like)*** *(triad)*
- ***Transition probability forward*** *(pair syllables 1 – 2)*
- ***Transition probability backward*** *(pair syllables 1 – 2)*
- ***Pointwise mutual information (PMI)*** *(pair syllables 1 – 2)*
- ***Variant of mutual information (MI-like)*** *(pair syllables 1 – 2)*
- ***Transition probability forward*** *(pair syllables 2 – 3)*
- ***Transition probability backward*** *(pair syllables 2 – 3)*
- ***Pointwise mutual information (PMI)*** *(pair syllables 2 – 3)*
- ***Variant of mutual information (MI-like)*** *(pair syllables 2 – 3)*
- ***1st syllable***
- ***Structure*** *(1st syllable)*
- ***Frequency*** *(1st syllable)*
- ***Percentage*** *(1st syllable)*
- ***Percentile rank*** *(1st syllable)*
- ***2nd syllable***
- ***Structure*** *(2nd syllable)*
- ***Frequency*** *(2nd syllable)*

- ***Percentage*** *(2nd syllable)*
- ***Percentile rank*** *(2nd syllable)*
- ***3rd syllable***
- ***Structure*** *(3rd syllable)*
- ***Frequency*** *(3rd syllable)*
- ***Percentage*** *(3rd syllable)*
- ***Percentile rank*** *(3rd syllable)*

Words/lemmas table

- ***Word/lemma***
- ***Frequency***
- ***Percentage***
- ***Percentile rank***

Pairs of words/lemmas table

- ***Pair of words/lemmas***
- ***Frequency*** *(pair)*
- ***Percentage*** *(pair)*
- ***Percentile rank*** *(pair)*
- ***Transition probability forward*** *(pair)*
- ***Transition probability backward*** *(pair)*
- ***Pointwise mutual information (PMI)*** *(pair)*
- ***Variant of mutual information (MI-like)*** *(pair)*
- ***1st word/lemma***
- ***Frequency*** *(1st word/lemma)*
- ***Percentage*** *(1st word/lemma)*
- ***Percentile rank*** *(1st word/lemma)*
- ***2nd word/lemma***
- ***Frequency*** *(2nd word/lemma)*
- ***Percentage*** *(2nd word/lemma)*
- ***Percentile rank*** *(2nd word/lemma)*

Triads of words/lemmas table

- ***Triad of syllables***
- ***Frequency*** *(triad)*
- ***Percentage*** *(triad)*
- ***Percentile rank*** *(triad)*
- ***Pointwise mutual information (PMI)*** *(triad)*
- ***Variant of mutual information (MI-like)*** *(triad)*
- ***Transition probability forward*** *(pair words/lemmas 1 – 2)*
- ***Transition probability backward*** *(pair words/lemmas 1 – 2)*
- ***Pointwise mutual information (PMI)*** *(pair words/lemmas 1 – 2)*
- ***Variant of mutual information (MI-like)*** *(pair words/lemmas 1 – 2)*
- ***Transition probability forward*** *(pair words/lemmas 2 – 3)*
- ***Transition probability backward*** *(pair words/lemmas 2 – 3)*
- ***Pointwise mutual information (PMI)*** *(pair words/lemmas 2 – 3)*
- ***Variant of mutual information (MI-like)*** *(pair words/lemmas 2 – 3)*
- ***1st word/lemma***
- ***Frequency*** *(1st word/lemma)*
- ***Percentage*** *(1st word/lemma)*
- ***Percentile rank*** *(1st word/lemma)*
- ***2nd word/lemma***
- ***Frequency*** *(2nd word/lemma)*
- ***Percentage*** *(2nd word/lemma)*
- ***Percentile rank*** *(2nd word/lemma)*
- ***3rd word/lemma***
- ***Frequency*** *(3rd word/lemma)*

- ***Percentage*** *(3rd word/lemma)*
- ***Percentile rank*** *(3rd word/lemma)*