# Enriched Orthographic Transcription: SPPAS convention.

*Brigitte Bigi – January 2017*

Before any kind of transcription, an automatic silence/speech detection must be performed: the units obtained this way are called IPUs (for Inter Pausal Units). Then, the orthographic transcription is performed on the basis of such IPUs. Notice that IPU boundaries must be manually verified.
In order to transcribe speech data, a convention must be defined. This convention has to take into account some methodological choice: using an automatic grapheme-to-phoneme system gives some constraints for the transcription. This document summarizes the convention supported by SPPAS.

## 1. Typographic rules

### *1.1 Abbreviations*

The convention doesn't allows any abbreviation, except if the developed form isn't normally used in written texts.

**Examples**:

madame Veil,
avant Jésus-Christ,
kilomètres à l'heure,
degrés Celsius

**But**:

etc

A specific dictionary is available in folder `resources/vocab/lang.repl`. In order to be used in the orthographic transcription, the new abbreviations must be appended in this file.

### *1.2 Numbers*

Numbers have to be written in letters only if they are not pronounced as expected.

**Examples:**

il est né en 2006
il est 3 heures

### *1.3 Punctuation*

SPPAS accepts the standard punctuation.

**Example**:

tu as relu : "le grand meaulnes" ?
j'ai relu "Le grand Meaulnes"
j'ai relu... Le grand Meaulnes !
j'ai relu le grand meaulnes

### *1.4 Acronyms, patronyms, toponyms*

The character '$' can be used to surround specific words like proper names. The convention also allows to use a TPS code, to append after a comma and before a slash:

- T for toponym;
- P for patronym;
- S for acronym.

The form is then: $Word, PTS/$.

**Examples**:

```
$Aix, T/$
$François$
$Senderens, P/$
```

*1.5 Spelled letters*

Spelled words are transcribed as a particular pronunciation (see section 2.2).

**Example:**

```
[ABC, abécé]
```

*1.5 Onomatopoeia*

They are included in the pronunciation dictionary. Take a look at it in `resources/dict/lang.dict`.

For example, in French, the typical back-channel onomatopoeia [m] produced by the hearer is noted as “mh” when it has one syllabus, and “mh mh” when it has two syllabus. Mutatis mutandis, for each language, depending on the corpus.

*1.6 Foreign words, regional words, etc.*

In that case, particular pronunciation convention must be used.

*1.7 Morphologic variants*

Graphic variants can be noted between <>, separated by commas.

**Example:**

```
<il chante, ils chantent>
```

## 2. Pronunciation notations

SPPAS only requires to mention unusual pronunciations. Standard pronunciations are all included in the pronunciation dictionary. This latter can be easily edited and modified if needed: `resources/dict/lang.dict.` For asian languages, the pinyin is accepted and can be mixed to character-based orthography. So, most of the time, this convention can be applied.

*2.1 Elisions*

When some phonemes are not pronounced by the speaker, surround the corresponding letters between parenthesis.

**Examples:**

Accepted: i(ls) sont v(e)nus / Recommended: ils sont venus
Accepted: i(l) y a / Recommended: il y a
Required: les arb(r)es

Notice that if a word is fully missing, it can't be surrounded by parenthesis:

**Examples:**

Forbidden: i(l) (y) a / Recommended: il y a
Forbidden: je (ne) sais pas / Recommended: je sais pas
Forbidden: je (ne) sais pas / Recommended: je {ne} sais pas

### *2.2 Specific pronunciations*

When a pronunciation cannot be expected at all from the standard orthography, the convention is as: *[standard, faked].*

**Examples:**

[je suis, chu]
[CB, cébé]
[copine, conpine]

Depending on the language, the pronunciation dictionary contains or not the most frequent reductions. In French, for example, it is not necessary to transcribe: [je sais, ché], [parce que, psk] because the dictionary already contains such pronunciations. It is then recommended to append a frequent specific pronunciation in the pronunciation dictionary instead of repeating it in the orthographic transcription of the corpus.

### *2.3 Word truncations*

They are noted by a final dash just after the final sound of the truncated word, and followed by a whitespace.

**Example:**

le pe- le petit
en pr- au collège

### *2.4 Liaisons*

Depending on the language, the pronunciation dictionary already includes or not the regular liaisons. Take a look at it to know about that. For French, all standard liaisons are already included, so they don't need to be mentioned in the orthographic transcription. However, unusual ones have to. The convention is to surround the missing letter by '=' symbol, without whitespace.

**Examples:**

trois amis (usual liaison not mentionned)
quatre =z= amis (unusual liaison)

If (for any reason) it is useful, missing standard liaisons can be mentioned using the # symbol surrounded by whitespace: trois # amis

## 3. Other phenomena

### *3.1 Reported speech*

Direct reported speech sequences can be annotated between symbols '§' surrounded by whitespace.

**Example:**

je lui ai dit § *je vois de quoi tu te plains §* ça lui a pas plu

### *3.2 Prosody break*

An unusual break in the prosody can be mentioned using '~' symbol surrounded by whitespace.

**Example:**

rien de (en)fin ~ ridicule quoi

### *3.3 Laughter*

Laugher items must be annotated with @ symbol surrounded by whitespace. Take a look at the SPPAS documentation to be sure SPPAS is supporting this convention for a given language.

When the speaker laughs while speaking, the convention allows to surround the word sequence between @@.

**Examples:**

C'est pas possible @
C'est pas @@ possible @@

### *3.4 Pauses*

Long pauses (i.e. silences) are automatically detected at a first stage, before transcribing. However, it is frequent that some shortest pauses occurs during speech. Such short perceptible pauses must be annotated with + symbol surrounded by whitespace.

**Example:**

je vois + tu es contente

### *3.5 Noises and incomprehensible sequences*

Long and short incomprehensible sequences and/or noises must be annotated by a star * surrounded by whitespace. Breathing, cough, sneeze, etc are all mentioned with this same symbol. Take a look at the SPPAS documentation to be sure SPPAS is supporting this convention for a given language.

## 4. Comments

Any comment of the transcriber can be added to the orthographic transcription by using braces. The only restriction is that the comment can't contain commas.

**Examples:**

ipu_172 {voix souriante} pas du tout

ipu_13 {tousse} *

ipu_203 * {inaudible} elle ét- c'était + ridicule