Transcription Convention

(release date: July 2023)

Brigitte Bigi

contact@sppas.org

In order to transcribe speech data, a convention must be defined. This convention has to take into account some methodological choices: using an automatic grapheme-to-phoneme system gives some constraints for the transcription.

This document describes the convention supported by SPPAS. It allows to perform manually an "Enriched Orthographic Transcription"...

1. Typographic rules

1.1 Specific abbreviations

The use of specific abbreviations is restricted to a given list. This list is stored into a specific dictionary and available in the folder *resources/vocab/*. These lists are ".repl" file extensions. In order to be used in the orthographic transcription, your custom abbreviations have to be appended into this file.

1.2 Numbers

SPPAS can convert numbers into their written form for most of the languages.

Example:

```
il est né en 2006, il est 3 heures.
```

But numbers have to be written in letters if they are not pronounced as expected:

Example:

```
il est né en dix-neuf-cent-nonente-dix, il est 3 heures.
```

1.3 Punctuation

SPPAS accepts the standard punctuation. Any of the following examples can be used.

Examples:

```
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 – not supported anymore.

Until SPPAS 4.11, the "$" character was used to surround proper names. The convention included to use a "TPS" code, to be appended after a comma and before a slash, with "T" for toponym, "P" for patronym and "S" for acronym.

Examples:

```
$Aix, T/$
$John Doe, P/$
$SPPAS, A/$
```

Any other "$" character was removed, which had side effects when transcribing English or any other language using the $ symbol to represent the word "dollars".
Then this is convention is no longer supported.

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 (see section 2.2).

1.7 Morphologic variants

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

Example:

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

2. Pronunciation notations

SPPAS only requires unusual pronunciations to be mentioned. 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 Broken words

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

Examples:

```
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 mentioned)
quatre =z= amis (unusual liaison)
```

If (for any reason) it is useful for you, a missing standard liaison 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. 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 Silences/Pauses

Long pauses (i.e. silences) are automatically detected at the first stage of the annotation workflow, before transcribing, so they can’t be added to the orthographic transcription with the ‘#’ character. However, the <cut> element (see section 5) could be used.

However, it is frequent that some shortest pauses occur 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
Hello {blink} how are you?
He{blink}llo, how are you?
```

5. Segments to be ignored – new in SPPAS 4.11

Any segment of speech that should be ignored can be mentionned with a “cut” tag. This tag must contain 2 attributes: “from” and “to” time values. Optionnally, it can indicates the time unit with the attribute “unit”, with either value “ms” for milliseconds or “sec” for seconds (the default). Any other attribute will be ignored.
SPPAS will split the segment into two elements: the one before the cut, and the one after. Several cut elements can be added into a same segment of speech.

Examples:

```
<cut from="5.68" to="5.74" />
<cut from="00:00:05.68" to="00:00:05.74" />
<cut unit="ms" from="5680" to="5740" />
<cut unit="ms" from="5680" to="5740" reason="pause" />
<cut unit="ms" from="5680" to="5740" why="laught" />
```

New in SPPAS-4.12:
An “event” tag can be used. It will be assigned as label of the cut interval.

Example:

```
<cut from="5.68" to="5.74" event=”+” />
```

6. Segments to be anonymized – new in SPPAS 4.12

Any segment of speech that should be buzzed can be surrounded by the “$” character, and it must not contain any space character – space can then be replaced by “-” or “_”.

Examples:

```
Here is $John_Doe$ accompanied by $D-J._Moon$ and…
This $fucking$ thing!
```