***Enriched orthographic transcription.***
***OTIM***

We developed these conventions in order to transcribe conversational data from the CID (Corpus of Interactional Data). They are made to allow the study of numerous interfaces between nearby domains of linguistics. We decided to use manual orthographic transcription, with more precisions about some particular pronunciations, and about some precise phenomena we wanted to study.

Standard orthographic transcription cannot describe phenomena we observe in oral spontaneous speech. Therefore we add precisions about atypical pronunciations (elisions, liaisons, particular pronunciations, etc.), disfluencies, laughers etc. We chose to enrich our orthographic transcription because we use an automatic processing of the data. Tools for automatic processing on standard French exist, but their results on spontaneous French data are not sure. And the gap between standard French and conversational corpus are unknown a priori. Consequently, we transcribe a lot of details, in order to make appropriate the use of these tools. The grapheme-phoneme converter we use generates a standard string of phonemes from an orthographic transcription. The aligner synchronizes this string of phonemes with the sound. A realistic string of phonemes improves the performance of the aligner. Transcribers have to give more information in the transcription, in order to make the string of phonemes nearer to the really-pronounced speech.

****

The conventions we present here have to take into account some methodological choices we did: segmentation, alignment and annotations should be automatic[1]. Respecting the constraints of the automatic grapheme-phoneme converter leads to better results with the software.

Our conventions are based on the GARS's ones. Because of the special kind of our data and the particular studies we wanted to drive on them, we modified these conventions. We needed to add some precisions about particular pronunciation, and some other details. This allows the apparition of oral specificities that couldn't appear with a standard orthographic transcription. The details we add on transcription avoid a lot of errors during automatic alignment, and so on it requires less manual corrections.

The one-hour-long dialogs are cut in smaller units. Before any kind of transcription (and annotation), an automatic segmentation of the interaction is made, based on the criteria of speech between 200ms pauses: the units we obtained this way are called IPU (for Inter Pausal Units). We recorded each speaker on a separate band. That's why we don't need to precise who is speaking, and when. We invite users to – as we did -- transcribe each speaker's speech on a separate tier.

********

[1] Transcription is made with software (Praat). Automatic alignment is made by LORIA and phonetic synthesis is made by Syntaix.

Our transcription is principally an orthographic transcription, without any punctuation. Because of the numerous automatic treatments made on this transcription, we add precisions about particular pronunciation (using the SAMPA conventions).

## 1. Typographic rules:

### *1.1. Abbreviations:*

NB: we distinguish abbreviation, truncation and elision.

Abbreviation: Mme pour Madame

Elision: le p(e)tit (see particular pronunciations)

Word truncations (3 kinds):

Involuntary truncations: `le li- le livre`

Apocope : `instit` for `institutrice`

Aphérèse: `blème` for `problème`

Our transcription doesn't note any abbreviation, except if the developed form isn't normally used in written texts. Elisions and word truncations are noted, we develop these points later in this document.

**Examples**:

- `madame Veil`
- `kilomètres à l'heure`
- `degrés Celsius`

**But** :

- `[etc, etsetera]` *(see particular pronunciations)*

### *1.2. Numbers :*

Numbers have to be written in letters.

**Examples:**

- `il est né en mille neuf cent dix`
- `il est trois heures`

### *1.3. Titles :*

Movies, books, newspapers titles are written between quotation marks, without a capital letter.

**Example**:

- `j'ai relu "le grand meaulnes"`

### *1.4. Acronyms, patronyms, toponyms*

We use a TPS code: T for toponym, P for patronym and S (from "sigle") for acronym. The form is: $Ortho,PTS/$. "Ortho" is the orthographic transcription.

Specific care must be put to acronyms whose pronunciation has the same form as spelled letters. (see spelled letters).

**Examples**:

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

### *1.5. Spelled letters:*

Spelled words are transcribed as particular pronunciation (between square brackets: capital letters, comma, and suggested pronunciation).

**Example:**

- `[ABC, abese]`

### *1.6. Onomatopoeia:*

We took a standard list of French onomatopoeia:

- `ah, aïe, areu, atchoum, badaboum, baf, bah, bam, bang, bé, bêêê, beurk, bien, bing, boum, broum, cataclop, clap clap, coa coa, cocorico, coin coin, crac, croa croa, cuicui, ding, ding deng dong, ding dong, dring, eh, eh ben, eh bien, euh, flic flac, flip flop, frou frou, glouglou, glou glou, groin groin, grr, hé, hep, hi han, hip hip hip hourra, houla, hourra, hum, mêêê, meuh, mh, miam, miam miam, miaou, oh., ouah, ouah ouah, ouais, ouf, ouh, paf, pan, patatras, pchhh, pchit, pff, pif-paf, pin pon, pioupiou, plouf, pof, pouet, pouet pouet, pouf, psst, ron ron, schlaf, snif, splaf, splatch, sss, tacatac, tagada, tchac, teuf teuf, tic tac, toc, tut tut, vlan, vroum, vrrr, wouah, zip.`

The typical back-channel onomatopoeia [m] produced by the hearer is noted as `mh` when it has one syllabus, and `mhm` when it has two syllabus. We distinguish `mhm` and `mh mh`.

We invite users to adapt the onomatopoeia conventions to the standard list of their language, depending on their corpus.

### *1.7. Other-languages words:*

In that case, we use particular pronunciation form in the TOE, and we can precise the language on another tier. We put in square brackets, and separated by comma, the standard transcription in the foreign language, and the speaker's pronunciation (in SAMPA alphabet).

**Example:**

- `[Christmas pudding, kRism9spudiN]`

### *1.8. Undeterminable morphologic variants*

Graphic variants are noted between braces, separated by commas.

**Example:**

- `{il chante, ils chantent}`

## 2. <u>Pronunciation notation :</u>

We only notate the most frequent, most important and most visible special pronunciations. When the transcriber doubts about the particular or standard pronunciation of a phrase, we give priority to the standard orthographic transcription.

### *2.1. Elisions:*

When some phonemes are not pronounced by the speaker, we put the corresponding letters between round brackets.

**Examples:**

- `i(l)s sont venus`
- `p(u)is`
- `t(u) as`

### *2.2. Particular pronunciations :*

When pronunciation cannot be visible by the round brackets, we use the particular pronunciation format: *[orthography, pronunciation]*

**Example:**

- `[je sais, Se]`
- `[c(el)ui, sHi]`

The grapheme-phoneme converter doesn't transcribe the "schwa" in the phonetic line. So the final schwa only can be found back thanks to the particular pronunciation notation.

### *2.3. Specific cases due to the grapheme-phoneme converter software:*

In French, "c-" would be /k/ before /a/, /o/ and /y/, if we don't precise a particular pronunciation. So we have to put it in square brackets. Similarly "c(e)" will also be understood as /k/, so we need to put round brackets in the square brackets. But we also keep the round brackets (in order to find automatically all the elisions in the corpus, as an example).

**Example:**

- `[c(e), s]` ➔ means that in the word "ce", only the consonant was pronounced, with a 'particular' pronunciation which is not /k/ but /s/.
- `[i(l)s ont, izo~]`

### *2.4. Atypical accords:*

We choose the standard orthographic transcription of the words as they have been said.

**Example:**

- `les conseils national` (instead of `"les conseils nationaux"`)

### *2.5. Unvolontary word truncations:*

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

**Example:**

- `le pe- le petit chien`

### *2.6. Liaisons*

The software which generates the phonemes also generates usual liaisons. This has three consequences: We don't notate usual liaisons, we notate unusual liaisons, we notate the absence of a usual liaison.

**Examples:**

- `trois amis` ➔ usual liaison
- `quatre=z=amis` ➔ unusual liaison
- `trois # amis` ➔ absence of usual liaison

### 3. Reported speech

Direct reported speech sequences are between symbols "§". It is preceded and followed by a blank. It has been showed that the beginning of direct reported speech is easier to find exactly, so the transcribers didn't have to note really precisely the end of the direct reported speech.

**Example:**

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

### 4. Incomprehensible sequences

Long and short incomprehensible sequences are always noted by one star (*).

### 5. Laughers

Laughers are noted with @. When the speaker laughs during his speech, when the word are said laughing, we put them between @@ and @@.

**Examples:**

- `c'est pas possible @` ➔ means that the speaker said the word, and then laughed.
- `c'est pas @@ possible @@` ➔ means that the speaker said "c'est pas", then laughed saying "possible"

### 6. Pauses

Long pauses (more than 200ms) are automatically detected. They are the boundaries of IPU. Short perceptible pauses are notated with "+".

**Example:**

- `je vois + tu es contente`

### 7. Non-linguistic events

They are noted as incomprehensible sequences, with a star. We suggest to precise what kind of event it was.

| | | |
|---|---|---|
| • Breathing | Respiration | `(r)` |
| • Puffing | Souffle | `(pf)` |
| • Noise by the mouth | Bruit de bouche | `(bb)` |
| • Cough | Toux | `(tx)` |
| • Sneeze | Eternuement | `(et)` |
| • Wistle | Sifflement | `(sif)` |