Orthographic transcription, tokenization, and lemmatization protocol for the project *SyllabO+*

Version of November 20, 2017

LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY

1. *Orthographic transcription*

1.1. General rules

- All transcriptions were executed by research assistants trained in linguistics and phonetics, under the supervision of Pascale Tremblay, Ph. D., and Johanna-Pascale Roy, Ph. D. The transcriptions were made from the audio recordings that constitute our corpus.
- The transcriptions are an accurate and non-normative reproduction of the speaker's vocabulary and syntax.[1]
- Prosody is not transcribed and neither are silences, laughs, onomatopoeia and other prosodic markers and punctuators. However, depending on the context, words with a specific function and meaning are transcribed. Examples: "*ah"*, "*euh"*, "*t'sais"*, "*ouais"*, "*pis"*, "*eille"*, "*oh"*, "*genre"*, "*comme"*, "*style"*, "*ok"*, etc.
- Non-speech sounds (breathing, coughing or other non-speech elements) are not transcribed, as well as hesitations, stuttering and other background noises made by the speaker or other people present during the recordings.[1]
- When transcriptions were obtained from a recording where the speaker was discussing in interaction with one or several other people, the distinct speech turns are kept in the transcription. Thus, only one speaker is transcribed, but his speech is not transcribed in one large unit, but rather in sections corresponding to his speech turns.
- All documents are transcribed orthographically, with proper spelling. Each term used by the speakers keeps its proper French spelling, regardless of pronunciation.[1]

1 Pichette, J-P. (2009). Protocole pour la transcription des documents de source orale en vue de l'édition. *Port Acadie : Revue interdisciplinaire en études acadiennes*, (16-17), 225-257. doi : 10.7202/045139ar

1.2. **Specific rules**

- **Transcriber's comments** :

  To ensure clarity of the transcription's content, comments or corrections from the transcriber will be in brackets [ ].

- **Verb conjugation**:

  Verb conjugations used by the speaker are transcribed as used, even if considered incorrect. They sometimes represent archaic forms. [1]

- **Syntax** :

  Speaker's syntax will be kept entirely as enunciated.[1] :

1.2..1. No word will be added to the transcription if not pronounced.[1]

1.2..2. The initial *ne* or *n'* of a negative statement will not be added, nor the conjunctive *que* if not used by the speaker.[1]

1.2..3. However, in certain cases, verbal ticks can be removed if they make reading unbearable, especially if their high frequency reduces them to a simple oral punctuation. It is important to discern and keep the ones which have a purpose.[1]

- **Anglicisms** :

  As with all words of foreign origin, anglicisms keep their regular spelling.[1]

- **Explanations** :

  Short dashes (−) are used to mark and separate from the rest of the narrative any explanation the speaker addresses to his listener(s).[1]

- **Numbers** :

  Dates keep their numbered form.

- Examples : "en 1980…", "dans les années 90…", "le 24 juin", etc.

  If an "et" is pronounced by the speaker, it is transcribed. (Ex : "soixante-et-douze")

All other numbers are transcribed in their lettered form.

- Examples : “dix-neuf ans”, “on était dix dans ma famille”, “trois-cents piastres”, “quatre-vingts heures”, “vingtième siècle”, “cinquante-trois virgule quatre pour cent”, etc.

- **Symbols** :

  Symbols are not used for transcription, but it is the lettered form that is used.

- Examples :
  - % = pour cent (“cinquante-trois virgule quatre pour cent”).
  - $ = dollar ou piastres (“cinq cents dollars”, “trois cents piastres”).
  - etc.

- **Incomprehensible words** :

  Incomprehensible words are transcribed by two consecutive sharp symbols ##.

- **Incompletely enunciated words** :

1.2..1. Sequences with an incomplete pronunciation are deleted from the text.

  - Example : “la ~~ma~~… la génération de mots”.

1.2..2. Exception for sequences where the word is still recognizable (for example, if only the last syllable is missing).

  - Examples :
    - Pronounced phrases : “Il y a quelqu’…”, “c’est dangere… imprudent”.
    - Transcribed phrases : “Il y a quelqu’un”, “c’est dangereux… imprudent”.

- **Abbreviated word** :

  An abbreviated word, which is not an incomplete pronunciation, but which is commonly used as a whole unit, is transcribed as pronounced.

- Examples : “chiro”, “psy”.

- **Elision** :

  A word which is elided by speaker is transcribed whole in written transcription.

  - Examples: *j'sais pas → je sais pas ; t'es → tu es* ; *t'sais → tu sais*
    - *Note : However, the "t'sais" used as a discursive marker keeps its elided form.* [2] [3] [4]

- **Measure units** :

  Measure units are written in their complete form, and not as an abbreviation.

  - Examples : "huit cents millisecondes", "vingt kilomètres", "huit grammes", etc.

- **Specific cases**:

  Certain specific cases, which are pronounced but are not necessarily listed in dictionaries, can be found in the document *Recensement orthographique* (attached at the end of this document). It contains each spelling chosen for the orthographic transcription as well as references accounting for these choices, when applicable.

  - *Examples : "t'sais", "OK", "mais que", etc.*

2 Pop, L. (2009). Quelles informations se pragmatisent*? Le cas des verbes plus ou moins marqueurs. *Revue roumaine de linguistique*, 54, 1-2 : 161-172.

3 Beaulieu-Masson, A., Charpentier, M., Lanciault, L. et Liu, X. (2007). Comme en français québécois. *Communication, lettres et sciences du langage*, 27-41.

4 Bolly, C. (2010). Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours : un no man's land entre syntaxe et discours? *Revue des linguistes de l'université Paris X Nanterre,* (62-63), 11-38. doi : 10.4000/linx.1356

2. *Tokenization*

Tokenization is the segmentation of text into lexical units. The word database was created from this tokenization process. This step was executed automatically by the *Treetagger* tool, integrated in a Python program. See http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ for documentation. Each word was recorded as is, with its marks for number (singular/plural), gender and conjugation. Note that all words were put in lowercase beforehand. For example, the sentence "J'ai mangé les raisins et une pomme." is divided this way: j / ai / mangé / les / raisins / et / une / pomme

Here are some details:

2.1. **Apostrophes** :

A syntagm containing an apostrophe is segmented in two distinct words by Treetagger.

- Examples : *j'ouvre (j / ouvre)*, *l'école (l / école)*, *j'ai (j / ai)*, *c'est (c / est)*, *s'est (s / est)*, *m'a (m / a)*, etc.

However, words containing an apostrophe as an integral part of the unit are not segmented by Treetagger.

- Examples : *aujourd'hui*, *d'accord*, *d'abord, quelqu'un*

2.2. **Hyphen** :

A two-word syntagm containing a hyphen are segmented in two distinct words by Treetagger (in particular verb-pronoun syntagms).

- Examples : *fait-il (fait / il)*, *voulez-vous (voulez / vous), etc.*

However, words containing a hyphen as an integral part of the unit are not segmented by Treetagger.

- Examples : *peut-être, arc-en-ciel, etc.*

2.3. **Single letters** :

Letters pronounced individually are counted as a single unit by Treetagger.

- Examples : “Donc le *n* est quand même assez bien”, “En trois *t* ?”, “Galaxy S”, « A, B ou C”, “pour *x* problème”, etc.

2.4. **Acronyms** :

Letters forming an acronym are considered as a single unit by Treetagger. The acronym is counted as a word.

- Examples : “PDF”, “RIL”, “FLQ”, “ONU”, “CSST”, “DEP”, etc.

3. *Lemmatization*

Lemmatization enables the grouping of lexical units by family, that is to assign a *lemma* (a canonical form) representing each word. The lemma database was created from this lemmatization process. This step was executed automatically by the *Treetagger* tool, integrated in a Python program. (following the tokenization process). See http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ for documentation. For example, the sentence “J’ai mangé les raisins et une pomme.” becomes : je / avoir / manger / le / raisin / et / un / pomme

Note : the process being automatic, the results can sometimes present certain anomalies, particularly in cases where the original sentence was ambiguous or if different grammatical categories (part-of-speech) were possible for a same word. Results are nevertheless adequate in a vast majority of cases and enable a good overview. Moreover, a post-treatment enabled the correction of many small anomalies.

He are more details :

3.1. **Verb conjugation :**

Les verbes conjugués sont remis à l’infinitif par Treetagger.

- Examples : viendrait → venir, est → être

3.2. **Gender and number (singular/plural) marks** :

When applicable, nouns and adjectives are reset to singular and masculine by Treetagger.

- Examples : étudiants → étudiant, jolie → joli

3.3. **Various notes:**

- In cases where Treetagger systematically defined an incorrect lemma, a post-treatment correction was executed. See Table 1.

- In cases where Treetagger defined always the same lemma, but the it should have varied according to contexts, the original word was counted instead. For example, the original word “des” systematically became the lemma “du”, when it could represent, according to context, either “de les” (i.e. “du” as its lemmatized form), either the plural of “un” or “une” (i.e. “un” as its lemmatized form). Given that the appropriate lemma was not systematically one or the other, the original word was counted instead of the lemma. See Table 1.
- If the selected lemma was "<unknown>", the original word was then counted. See Table 1.
- In cases where Treetagger found an ambiguity and proposed two options, the selected lemma was the one corresponding to the correct choice for the majority of cases. See Table 1.
- In many cases, Treetagger defined “Verbe, participe passé” (Verb, past participle) as the grammatical category (part-of-speech) and thus defined the verb (infinitive form) as the lemma. Note that this often included words that could rather be considered as adjectives. The treatment executed by Treetagger was however kept untouched; it is simply important to keep in mind this particularity.
    - Examples:
        - “je suis passé par là” → “passé” = “passer” (lemma)

            But also:
        - “c'était la semaine passée” → “passée” = “passer” (lemma)
        - “il a ouvert une porte” → “ouvert” = “ouvrir” (lemma)

            But also:
        - “une fenêtre ouverte sur le monde” → “ouverte” = “ouvrir” (lemma)

Table 1.

Treetagger Lemma	**Selected Lemma**
-ci	ci
<unknown>	*mot original*
aciduler	acidulé
acteur|actrice	acteur
ado|ados	ado
ailler	aller
anglais|anglaise	anglais
angler	anglais
anser	anse
argenter	argenté
argentin|argentine	argentine
arrivé|arrivée	arrivée
articler	article
artistiquer	artistique
bai	baie
bai|baie	baie
banquer	banque
barbeler	barbelé
bégayer|bégayer	bégayer
bienvenir	bienvenue
bienvenu|bienvenue	bienvenue
blond|blonde	blonde
bordeau|bordeaux	bordeaux
bouiller|bouillir	bouillir
bouter	bouton
boutonnier|boutonnière	boutonnière
bri|bris	bris
bruir|bruire	bruit
brun|brune	brun
canadien|canadienne	canadien
carrer	carré
cela	*mot original*
chimiquer	chimique
chiner	chine
chip|chips	chips
cinématographier	cinématographie
clémentine|clémentines	clémentine
commun|communs	commun
convenir|convier	convenir
coordonnée|coordonnées	coordonnée
cour|cours	cours
croire|croître	cru
cuisinier|cuisinière	cuisinier
cycler	cycle
damer	dame
déblayer|déblayer	déblayer
dégueulasser	dégueulasse
demi|demie	demi
dépeigner|dépeindre	dépeindre

directeur|directrice	directeur
docteure	docteur
doctriner	doctrine
douteur|douteux	douteux
droit|droite	droite
du	*mot original, si “des”*
duplicater	duplicate
échec|échecs	échec
effrayer|effrayer	effrayer
essayer|essayer	essayer
être|êtres	être
ficher|ficher	ficher
fil|fils	fils
foi|fois	fois
folle|fou	fou
fond|fonds	fonds
fonder|fondre	fondre
force|forces	force
frai|frais	frais
gille	gilles
graisseur|graisseux	graisseux
gros|grosse	gros
guide|guides	guide
hôte|hôtesse	hôte
illimiter	illimité
indien|indienne	indien
intimer	intime
jacque|jacques	jacques
jardinier|jardinière	jardinière
journal|journau	journal
la|le	le
lac|lacs	lac
laisse|laisses	laisser
laurer	*mot original*
lettrer	lettre
li|lis	lire
limbe|limbes	limbes
lire|liser	lire
liser	*mot original*
logo|logos	logo
lunette|lunettes	lunette
maître|maîtresse	maître
malais|malaise	malaise
malter	malte
marger	marge
matériau|matériaux	matériau
maximer	maxime
mécher	méchant
méditerrané	méditerranée
moi|mois	mois
monnayer|monnayer	monnayer
moteur|motrice	moteur
orphelin|orpheline	orphelin

ouais	oui
ouater	ouaté
ouvrer|ouvrir	ouvrir
pâque|pâques	pâques
parage|parages	parage
parent|parents	parent
patent	patente
patron|patronne	patron
payer|payer	payer
pétroler	pétrole
piser	pise
plaire|pleuvoir	plaire
pleuroter	pleurote
poirer	poire
policer	police
poste|postes	poste
premier|première	premier
recouvrer|recouvrir	recouvrir
réessayer|réessayer	réessayer
règle|règles	règle
remblayer|remblayer	remblayer
ressortir|ressortir	ressortir
retraiter	retraité
rouget|rougette	rougette
roulotter	roulotte
routiner	routine
second|seconde	seconde
sen|sens	sens
sommer|être	être
souffleur|souffleuse	souffleuse
spiral|spirale	spirale
suite|suites	suite
suivre|être	être
taler	talent
tau|taux	taux
tourbier|tourbière	tourbière
travailleur|travailleuse	travailleur
veilleur|veilleuse	veilleuse
venu|venue	venu
ver|vers	ver
veuf|veuve	veuf
vitaminer	vitamine