# Protocole de transcription orthographique,

# de tokenisation et de lemmatisation du projet *SyllabO+*

Version du 20 novembre 2017

LABORATOIRE DES NEUROSCIENCES
DE LA PAROLE ET DE L'AUDITION

SPEECH AND HEARING
NEUROSCIENCE LABORATORY

# 1. *Transcription orthographique*

## 1.1. **Règles générales**

- La transcription a été effectuée par des auxiliaires de recherche ayant obtenu une formation en linguistique et en phonétique, sous la supervision de Pascale Tremblay, Ph.D. et de Johanna-Pascale Roy, Ph.D. La transcription a été faite à partir de l'écoute des enregistrements audio.
- La transcription est une reproduction intégrale (non normative) du vocabulaire et de la syntaxe du locuteur, par respect et fidélité à l'égard de sa langue.[1]
- La prosodie n'a pas été retranscrite, de même que les silences, rires, onomatopées et autres marqueurs prosodiques ou ponctuants. Cependant, selon le contexte, certains mots avec une fonction et un sens propres sont transcrits. Exemples : « ah », « euh », « t'sais », « ouais », « pis », « eille », « oh », « genre », « comme », « style », « ok », etc.
- Les bruits ambiants (respiration, toux ou autres éléments de non parole) ne sont pas retranscrits, de même que les hésitations, bégaiements et autres bruits parasites provenant du locuteur ou de son auditoire.[1]
- Lorsque la transcription d'un locuteur est issue d'un enregistrement où le locuteur parle en interaction avec une ou plusieurs autres personnes, les tours de paroles distincts sont conservés dans la transcription. Ainsi, un seul locuteur est transcrit, mais son discours n'est pas en un seul bloc, il est transcrit en sections correspondant à ses tours de paroles.
- Tous les documents sont transcrits de manière orthographique, avec rétablissement de l'orthographe. « Chaque terme utilisé par le [locuteur] conservera sa graphie en français courant, quelle que soit la prononciation qui lui sera attribuée. » [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. **Règles spécifiques**

- **Commentaires du transcripteur** :

  Pour plus de clarté sur le contenu de la transcription, les commentaires ou corrections du transcripteur sont mis entre crochets [ ].

- **Désinences verbales** :

  Les désinences verbales employées par le locuteur sont reproduites telles quelles.

  « Ces conjugaisons, considérées de nos jours irrégulières ou fautives, maintiennent parfois vivantes des formes archaïques. » [1]

- **Syntaxe** :

  « La syntaxe du locuteur sera intégralement respectée » [1] :

  1.2..1. « Aucun mot nouveau ne sera ajouté sans raison suffisante. » [1]

  1.2..2. « On ne devra pas ajouter la partie initiale *ne* ou *n'* de la négation ou le *que* de la conjonction si le locuteur ne l'emploie pas. »[1]

  1.2..3. « Cependant, on pourra parfois retrancher […] les chevilles et tics verbaux du locuteur qui deviendraient ahurissants ou insupportables à la lecture, notamment lorsque leur fréquence élevée les réduit à une simple ponctuation orale. »[1]

  « On veillera à bien discerner et à conserver ceux qui gardent vraiment leur raison d'être. » [1]

- **Anglicismes** :

  « Comme les autres mots d'origine étrangère, les anglicismes garderont leur graphie habituelle. »[1]

- **Didascalies** :

  « Les tirets courts (−) serviront à détacher du récit les didascalies et autres explications que le locuteur adresse à ses auditeurs. » [1]

- **Chiffres** :

  Les dates conservent leur forme chiffrée.

  - Exemples : « en 1980… », « dans les années 90… », « le 24 juin », etc.

  Si un « et » est prononcé par le locuteur, il sera transcrit (ex : soixante-et-douze)

  Tous les autres chiffres sont écrits sous leur forme lettrée.

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

- **Symboles** :

  Les symboles ne sont pas utilisés, mais c'est leur forme lettrée qui est utilisée.

  - Exemples :
    - % = pour cent (« cinquante-trois virgule quatre pour cent »).
    - $ = dollar ou piastres (« cinq cents dollars », « trois cents piastres »).
    - etc.

- **Mots incompréhensibles** :

  Les mots incompréhensibles seront transcrits par deux symboles de dièse consécutifs ##.

- **Mots incomplets dans leur prononciation** :

1.2..1. Les séquences dont la prononciation n'est pas complète seront supprimées du texte.

- Exemple : « la ~~ma~~… la génération de mots ».

1.2..2. Exception pour les séquences dont on reconnaît facilement le mot (comme s'il ne manquait que la dernière syllabe, par exemple).

- Exemples :
  - Phrases prononcées : « Il y a quelqu'… », « c'est dangere… imprudent ».

- Phrases transcrites : « Il y a quelqu'un », « c'est dangereux… imprudent ».

- **Mots abrégés** :

  Un mot abrégé, qui n'est pas une prononciation incomplète, mais qui est utilisé tel quel dans l'usage courant, sera transcrit comme prononcé.

  - Exemples : « chiro », « psy ».

- **Élision**:

  Un mot élidé à l'oral est transcrit en entier à l'écrit.

  - Exemples: *j'sais pas □ je sais pas ; t'es □ tu es* ; *t'sais □ tu sais*
    - *Note : Le « t'sais » utilisé comme marqueur discursif conserve toutefois sa forme élidée* [2] [3] [4]

- **Unités de mesure** :

  Les unités de mesure seront écrites sous leur forme lettrée complète, et non sous leur forme abrégée.

  - Exemples : « huit cents millisecondes », « vingt kilomètres », « huit grammes », etc.

- **Cas particuliers** :

  Certains cas particuliers prononcés, mais n'étant pas répertoriés dans les dictionnaires, se retrouvent dans le document *Recensement orthographique* (ci-joint à la fin du document). La graphie retenue pour la transcription orthographique et les références expliquant le choix sont présentes.

  - *Exemples : « 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. *Tokenisation*

La tokenisation est la segmentation du texte continu en unités lexicales. C'est du processus de tokenisation qu'est issue la base de données de mots. Cette étape a été effectuée de manière automatique par l'outil *Treetagger*, intégré dans un script Python. Voir http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ pour la documentation. Chaque mot a été comptabilisé tel quel, avec ses marques de genre, de nombre et de conjugaison. À noter que tous les mots avaient d'abord été réduits en lettres minuscules. Par exemple, la phrase « J'ai mangé les raisins et une pomme. » est divisée ainsi : j / ai / mangé / les / raisins / et / une / pomme

Voici quelques précisions:

### 2.1. **Apostrophes** :

Un syntagme contenant une apostrophe est segmenté en deux mots distincts par Treetagger.

- Exemples : *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.

Par contre, les mots qui contiennent une apostrophe comme partie intégrante de l'unité ne sont pas segmentés.

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

### 2.2. **Traits d'union** :

Un syntagme de deux mots contenant un trait d'union est segmenté en deux mots distincts par Treetagger (notamment les suites de verbes et pronoms).

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

Par contre, les mots qui contiennent un trait d'union comme partie intégrante de l'unité ne sont pas segmentés.

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

2.3. **Lettres seules** :

Les lettres prononcées seules sont comptées comme une unité par Treetagger.

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

2.4. **Acronymes** :

Les lettres formant un acronyme sont considérées comme une seule unité. L'acronyme est donc compté comme un mot.

- Exemples : « PDF », « RIL », « FLQ », « ONU », « CSST », « DEP », etc.

## *3. Lemmatisation*

La lemmatisation permet de regrouper les unités lexicales par famille, c'est-à-dire d'assigner un *lemme* (une forme canonique) qui représente chaque mot. C'est du processus de lemmatisation qu'est issue la base de données de lemmes. Cette étape a été effectuée de manière automatique par l'outil *Treetagger*, intégré dans un script Python (à la suite de la tokenisation). Voir http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ pour la documentation. Par exemple, la phrase « J'ai mangé les raisins et une pomme. » devient alors : je / avoir / manger / le / raisin / et / un / pomme

À noter : le processus étant automatique, les résultats obtenus peuvent parfois contenir certaines anomalies, notamment dans les cas où la phrase originale était ambiguë ou s'il y avait plusieurs catégories grammaticales possibles pour un mot. Les résultats sont tout de même adéquats dans la majorité des cas et permettent d'avoir une bonne vue d'ensemble. De plus, certaines rectifications post-traitement ont permis de faire disparaître plusieurs petites anomalies.

Voici quelques précisions :

3.1. **Conjugaisons verbales :**

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

- Exemples : viendrait → venir, est → être

3.2. **Marques de genre et de nombre** :

Lorsque applicable, les noms et les adjectifs sont remis au singulier et au masculin par Treetagger.

- Exemples : étudiants → étudiant, jolie → joli

3.3. **Notes diverses :**

- Dans les cas où Treetagger déterminait systématiquement un lemme erroné, une rectification était effectuée (post-traitement). Voir Tableau 1.
- Dans les cas où Treetagger déterminait toujours un même lemme, mais qui aurait dû varier selon les contextes, le mot original était alors comptabilisé. Par exemple, le mot original « des » devenait systématiquement le lemme « du », alors qu'il pouvait représenter soit « de les » (donc effectivement « du » sous sa forme lemmatisée), soit le pluriel de « un » ou « une » (donc « un » sous sa forme lemmatisée). Comme ce n'était pas systématiquement l'un ou l'autre choix qui était approprié, c'est le mot original qui a été comptabilisé plutôt que le lemme. Voir Tableau 1.
- Si le lemme attribué était "<unknown>", c'est le mot original qui était alors comptabilisé. Voir Tableau 1.
- Dans les cas où Treetagger trouvait une ambiguïté et proposait deux options, le lemme sélectionné était celui qui correspondait au choix correct dans la majorité des cas. Voir Tableau 1.
- Dans de nombreux cas, Treetagger a déterminé « Verbe, participe passé » comme catégorie grammaticale (et détermine donc le verbe à l'infinitif comme lemme). À noter que cela inclut souvent certains mots qui pourraient plutôt être considérés comme adjectifs. Le traitement effectué par Treetagger a été conservé tel quel, donc il est important de garder en tête cette particularité.
    - Exemples:
        - « je suis passé par là » → « passé » = « passer » (lemme)

          Mais aussi:
        - « c'était la semaine passée » → « passée » = « passer » (lemme)
        - « il a ouvert une porte » → « ouvert » = « ouvrir » (lemme)

Mais aussi :

- « une fenêtre ouverte sur le monde » → « ouverte » = « ouvrir » (lemme)

Tableau 1.

| **Lemme Treetagger** | **Lemme sélectionné** |
|---|---|
| -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 |