Tutoriel : transcription orthographique

Anne DISTER

La transcription orthographique des données du corpus Rhapsodie suit strictement les quatre grands principes suivants :

1. **Adoption stricte de l'orthographe standard.** On n'utilisera donc aucun des "trucages orthographiques" qui visent à calquer la prononciation.

On écrira *parce que* et non **pasque, ils vont* et non **i vont, ils ont* et non **i-z-ont,* etc.

De la même manière, on ne pratiquera aucune élision. On transcrira donc *tu as* et non **t'as, peut-être* et non **p'têt'*, etc.

Le principe général de transcription dans Rhapsodie est de matérialiser la présence ou l'absence des morphèmes, mais non de s'intéresser à la forme particulière qu'ils peuvent prendre. C'est donc la forme la plus conventionnelle des morphèmes qui est transcrite, et les morphèmes non prononcés ne sont pas transcrits.

2. **Absence de ponctuation.** Les transcriptions ne sont pas ponctuées, et cela afin de ne pas orienter les analyses syntaxiques ultérieures. Aucun signe de ponctuation n'est utilisé : ni le point d'interrogation dans le cas des questions, ni le point dans les sigles. Les prises de parole des locuteurs ne commencent donc pas non plus par une majuscule, le concept de phrase étant abandonné. La majuscule est utilisée uniquement pour les sigles, les acronymes et les noms propres (cf. ci-dessous).

3. **Aucune notation de pause silencieuse.** Les pauses silencieuses ne sont pas notées dans la transcription orthographique, ce paramètre relevant du versant prosodique.

4. **Prise en compte de l'oralité des données**, avec notation minutieuse des phénomènes suivants :

- la pause dite pleine : *euh*

- les répétitions de mots (p. ex. : *le le papier*)

- les auto-corrections (p. ex. : *le la carte*)

- les amorces de morphemes (p. ex. : *on s'ad~ on s'adapte*)

- les interjections et onomatopées (cf. liste non exhaustive ci-dessous)

- les chevauchements de parole

Le tableau ci-dessous synthétise les conventions uilisées :

Phénomène	Marque	exemple
Acronymes	1re lettre en majuscule	Il travaille à la Nasa

Alternance de code[1]	$ $	alors $ yes we can $ qu'il me dit
Amorces de morphèmes	~	une piste d'atterrissage d'hélicop~ pour héli~ hélicoptère
Apocopes et aphérèses	aucune marque	Steph de Monac, ricain
Nombres	écrits en toutes lettres	septante-deux, soixante-douze
Noms propres	1re lettre en majuscule	George Clooney
Passages inaudibles	***	
Sigles	tout en majuscule, sans espace	maintenant il pointe à l'ANPE
Titres (livres, films, etc.)	" "	il a siffloté "Tea for two"

Abréviations graphiques

Aucune abréviation graphique — procédé usuel à l'écrit — n'est employée quand un mot est prononcé dans son intégralité. Tous les termes sont transcrits en entier.

Accents graphiques

Tous les mots transcrits sont accentués, y compris sur les lettres majuscules.

Accords non standards

Les accords non standards audibles ne sont pas corrigés : ils sont transcrits à l'aide des morphèmes usuellement utilisés. Plus généralement, tout ce qui peut apparaitre comme une « faute » de grammaire est transcrit verbatim, sans correction. On n'ajoutera aucun *(sic)* dans ce cas.

il y a

La tournure *il y a* peut être prononcée de 4 manières différentes: [ilia], [ilja], [ija] ou [ja]. On a en outre un continuum entre ces différentes prononciations qui, souvent, ne peuvent aisément être distinguées. La règle est de transcrire systématiquement *il y a*.

Multitranscriptions

Certaines conventions proposent des multitranscriptions en cas de doute, qu'il s'agisse d'un problème d'écoute ou d'un choix grammatical, au niveau des accords par exemple (cf. : *ses frères et sœurs* vs *ses frère et sœurs* vs *ses frères et sœur* vs *ses frère et sœur*). Dans Rhapsodie, le transcripteur choisit la solution qui lui parait la plus plausible selon le contexte.

Le *ne* de négation

[1] L'alternance de code concerne les passages longs, et non les emprunts. On ne considèrera pas comme une alternance de code *c'est un has been*.

L'adverbe de négation *ne* n'est pas audible lorsqu'il est précédé de *on* et suivi d'une voyelle ou d'un *h* muet. Par exemple, quand on entend [o~napAlta~], on ne peut pas dire si le locuteur a construit l'énoncé *on n'a pas le temps* ou bien *on a pas le temps*.

Dans ce cas, ce possible *n'* de négation n'est *jamais* transcrit, et ce même si le même locuteur prononce explicitement des *ne* de négation dans d'autres énoncés.

Particules de l'oral

Les « particules de l'oral » (interjections, onomatopées, etc.) sont transcrites de manière normalisée. Le tableau ci-dessous recense les graphies de quelques particules courantes. On se reportera au *Dictionnaire des onomatopées* (Enckell et Rézeau, 2005) pour une liste plus complète.

ah aïe bah bé	ben eh euh hein	Hum mais enfin mh moui mouais	oh oh la la ouille pf

Les variantes mineures d'onomatopées ne sont pas distinguées. Par exemple, c'est toujours *pf* qui est transcrit, jamais *pff*, *pfff* , ou *pffff*.

La particule *mh* se distingue de *hum* de la manière suivante : *mh* correspond à un acquiescement du locuteur (parfois répété : *mh mh*), *hum* est utilisé dans les autres cas.

Phénomènes phonétiques et prosodiques

Les phénomènes phonétiques et prosodiques (prononciations particulières, élisions, allongements vocaliques, liaisons, reprises de souffle, pauses, intonation) ne sont pas transcrits, de même que les prononciations non standards, les liaisons erronées, etc. (Par ex. [yiz9Ro] pour *huit euros*).

Format des fichiers de transcription orthographique

La transcription orthographique se fait directement sous le logiciel Praat (http://www.fon.hum.uva.nl/praat/) en alignant le texte au son. Chaque locuteur se voit attribuer une tire dans laquelle toutes ses paroles sont transcrites. Les intervalles dans lesquels figurent les paroles transcrites ne doivent en aucun cas être considérés comme des unités théoriques, de quel que niveau que ce soit (ni syntaxique, ni prosodique). Il s'agit simplement de segments, de longueur relativement brève, qui permettront de faciliter les étapes ultérieures de la phonétisation faite avec Easyalign (cf. ***).

Si la transcription se fait sous Praat, une version de la transcription orthographique est aussi disponible sous la forme d'un texte suivi (ou *texte continu*) dans un fichier de type « .txt » : ce texte offre une vue fusionnée et linéarisée des tires, débarrassée des informations d'alignement. C'est cette forme qui sert de base pour les annotations syntaxiques.