CONVENTIONS DE TRANSCRIPTION EN VUE D'UN ALIGNEMENT TEXTE-SON AVEC TRANSCRIBER

Virginie André, Emmanuelle Canut, Jeanne-Marie Debaisieux, Christophe Benzitoun

Préambule

La transcription d'un enregistrement sonore tente d'être la plus fidèle possible aux paroles prononcées par les locuteurs. Néanmoins, elle ne peut pas refléter l'enregistrement avec une fidélité parfaite : la prononciation d'un même terme, par un enfant comme par un adulte, est variable et peut parfois être difficilement identifiable. Les éléments transcrits sur lesquels subsistera un doute ne seront ni interprétés ni analysés, seules des hypothèses pourront éventuellement être formulées. Les transcriptions TCOF tentent également d'être les plus lisibles possibles afin de permettre différents types d'analyse. C'est pourquoi nous avons choisi de respecter l'orthographe standard sans aménagement.
Le but de notre travail sur corpus est une étude approfondie du fonctionnement du français parlé, des points de vue syntaxique, sémantique, pragmatique ou encore interactionnel.

Tours de parole

Chaque tour de parole fait l'objet d'une identification du locuteur dans TRANSCRIBER. Cette identification est représentée sur l'écran par un encadré. *Pour les corpus adulte-enfant, l'adulte est identifié par « adulte » et l'enfant par son prénom. Pour les corpus entre adultes les locuteurs sont identifiés par « L » et sont numérotés dans l'ordre de leur prise de parole : L1, L2, etc*. On obtient la présentation suivante :

Corpus adulte-enfant :

adulte
- ta maman vient te chercher

David
- non

adulte
- ah bon

Corpus entre adultes :

L1
- et concrètement qu'est ce qu'il a fait pour euh

L2
- euh plein de choses euh en fait euh grâce à lui euh j'ai pu rencontrer Jackie Chan euh la semaine de mon arrivée

L1
- hum hum

Pour les corpus adulte-enfant, les énoncés doivent être numérotés mais la numérotation n'est pas à réaliser au moment de la transcription. Tous les énoncés seront numérotés automatiquement dans des balises « commentaire » après la première phase de transcription et avant la phase de vérification grâce à un programme informatique.

Principes généraux de transcription

1. Transcrire tout ce qui est dit par l'adulte et par l'enfant, y compris les hésitations et les répétitions, selon l'orthographe usuelle. Ne pas ajouter d'éléments non verbalisés. Par exemple, ne pas écrire « ne » lorsque cette partie de la négation n'est pas réalisée.

2. Ne pas ponctuer les énoncés.
   Ne pas employer les majuscules sauf pour les noms propres. Mettre des guillemets droits pour les titres, avec majuscule sur le premier mot (ex : "La belle et la bête").
   Exception pour les corpus adulte-enfant : indiquer « ? » pour exprimer l'intonation interrogative.

3. Ne pas anonymiser le corpus (indiquer tous les noms propres, chiffres etc. entendus, même ceux permettant l'identification de locuteurs). Le balisage des éléments susceptibles d'être anonymisés et l'anonymisation (dans le son et le texte) se fera ultérieurement selon un codage spécifique.

Orthographe et prononciation : spécificités

- Les nombres doivent respecter les normes habituelles de l'écrit. Tous les nombres doivent être écrits en lettres, sauf les années.

- Les phatiques et les onomatopées sont codifiés et doivent être transcrits selon l'orthographe fournie :

ah, aïe, areu, atchoum, badaboum, baf, bah, bam, bang, bé, bêêê, beurk, ben, bing, bon, boum, broum, cataclop, clap clap, coa coa, cocorico, coin coin, crac, croa croa, cuicui, ding, ding deng dong, ding dong, dring, hé, hé 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, miam, miam miam, miaou, oh, O.K., 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.

- Les sigles sont ponctués quand on lit les lettres isolément (S.N.C.F.), non ponctués lorsqu'il s'agit d'acronyme (CROUS). Le sens des sigles peut être précisé lors de la première apparition avec une balise TRANSCRIBER : A.E.E. *{sigle=Agence Européenne de l'Environnement}*

- Lorsque l'orthographe est incertaine (cas de certains noms de marques, toponymes, etc.) on pourra utiliser une orthographe approximative si elle est plausible et le signaler avec des balises TRANSCRIBER, soit à la suite du mot isolé : un certain Dupont+*[lex=?]* ; soit en englobant la séquence : *[lex=?-]*Sous la Voivre*[-lex=?]*.

- Les accords non standards sont suivis de la balise (commentaire) TRANSCRIBER {sic} : « tu as vu des chevals {sic} », « ils croivent {sic} que c'est vrai ».

- Respecter les règles d'accord sauf si on a une réalisation phonique particulière. Par exemple, « on est **parti** avec maman » mais « on s'est **mises** à dormir ».

- Quand il y a hésitation entre plusieurs possibilités pour la transcription, noter les mots entre barres obliques, séparées par une virgule : /ça, chat/ ou /va, vois/.

- Quand on a des mots indistincts (incompréhensibles, inaudibles, inconnus), mettre * si cela ne s'applique qu'à une syllabe, *** si cela s'applique à plusieurs syllabes. *On pourra éventuellement ajouter une balise prononciation : c'est une * [pron=pul ?].*

- Les liaisons particulières sont indiquées entre deux signes = : « le =n= ours », « donne-moi =z= en ».

- On utilisera les parenthèses pour les variantes morphologiques indécidables (non réalisées à l'oral) : « il(s) disai(en)t… », « on (n') est pas là ».

- Ne pas rétablir les élisions non réalisées : « parce que il est pas là ».

- Dans le cas où sont verbalisés des mots en langue étrangère, transcrire selon la norme de la langue d'origine et insérer une balise (langue) de TRANSCRIBER. On obtient par exemple la configuration : speed+*[lang=anglais]*

SPECIFICITES POUR LES CORPUS ADULTE-ENFANT

- Dans les corpus adulte-enfant, on pourra utiliser {sic} quand il y a reprise par l'adulte d'un mot « tronqué » verbalisé préalablement par l'enfant :

enfant

- non c'est un éléphant *[pron=efa~]*

adulte

- c'est un oui c'est un éphant *{sic}* c'est un éléphant

- Dans les corpus adulte-enfant, quand certains éléments (syllabes) ne sont pas réalisés à l'intérieur d'un mot ou quand le mot est parfaitement reconnaissable mais que ponctuellement il est verbalisé avec une prononciation un peu différente (inversion de syllabes par ex) : écrire le mot correctement orthographié puis dans TRANSCRIBER insérer une balise de prononciation et écrire la prononciation exacte entendue avec l'alphabet SAMPA. Sur l'écran de transcriber, on obtient la configuration suivante : il frappe [pron=ifap] ; elle court [pron=Ecu] ; parce que [pron=pak2] ; tu as [pron=ta], je sais [pron=SE], spectacle [pron=pEstakl], chat [pron=sa], pomme [pron=pEm], etc.

En revanche, ne pas mettre de balises de prononciation pour la prononciation élidée du schwa (mots comportant un « e » susceptible d'être prononcé et qui ne l'est pas) et conserver uniquement la forme orthographique usuelle : transcrire petit pour [pti].

Dans le cas d'une prononciation particulière récurrente dans le corpus, par exemple l'enfant prononce [z] tous les [Z] (« ze » pour « je »), elle ne sera pas reportée dans la transcription mais indiquée en commentaire dans la fiche signalétique.

- Si ce que dit un enfant n'est pas compréhensible ou s'éloigne assez d'une prononciation standard, ne pas écrire de mots ou segments en orthographe standard mais utiliser directement dans TRANSCRIBER une balise de prononciation [pron=pap2mut] suivie d'un commentaire {interprétation=pamplemousse}.
Dans le cas d'un doute ou d'une impossibilité d'interprétation, mettre un point d'interrogation dans l'accolade : [pron=iai] {interprétation=parti ?} ; [pron=emusEbele] {interprétation=la mouche s'est envolée ?}.

Dans les énoncés de l'enfant, quand il y a un doute sur l'interprétation du morphème réalisé, transcrire en alphabet SAMPA dans une balise de prononciation et mettre à la suite, dans une balise commentaire, les deux interprétations possibles : [pron=ki] {/qui, qu'il/} ; [pron=i] {/il, qui/}

Phénomènes propres à l'oral

Chevauchements de paroles
=> indiquer les chevauchements entre les locuteurs avec la balise « locuteur superposé » dans TRANSCRIBER. Sur l'écran on obtient la configuration suivante :

L1

- je pense que c'est

L1 + L2

- 1- toujours lié à l'ordinateur
  2- et euh sou- souvent souvent

L2

- on y pense pas

Dans ce cas, 1- correspond à L1 et 2- à L2 et les segement 1 et 2 sont prononcés en même temps.

- Noter les amorces de mots par un tiret (collé) : « il a il a p- il a pris »
- Les pauses sont notées par une croix (espace avant et après) : +

- Les pauses très longues (silence) sont notées /// et peuvent être accompagnées d'une balise (commentaire) de TRANSCRIBER : /// {l'enfant regarde attentivement l'image} ; /// {l'enfant va chercher un jouet dans sa chambre} ; /// {L1 se sert un verre d'eau}.

- Préciser les détails de la situation (rires, bruits, etc.) avec une balise (commentaire) de TRANSCRIBER. Si l'événement ou le « bruit » est répertorié dans TRANSCRIBER, insérer cette balise (ex : [rire]). Si l'événement ou le « bruit » n'est pas répertorié, le décrire : {bruit d'une sirène dans la rue} ; {éternuement de l'adulte} ; {l'enfant saute sur son lit} ; {une porte qui claque}.
  Si un événement est récurrent (bruits de fond, etc.), il ne sera pas reporté dans la transcription mais indiqué en commentaire dans la fiche signalétique.

- Les parties enregistrées mais non transcrites sont notées ### suivi d'une balise (commentaire) de TRANSCRIBER : ### {la mère entre dans la chambre et pose une question à l'enfant}.

- Les coupures dans l'enregistrement sont indiquées par $$$ suivi d'une balise (commentaire) de TRANSCRIBER : $$$ {l'enfant éteint le magnétophone par inadvertance} ; $$$ {sonnerie du téléphone : l'adulte arrête l'enregistrement pour répondre} ; $$$ {partie confidentielle, les locuteurs ont demandé l'arrêt du magnétophone}.

RECAPITULATIF DES SYMBOLES DE TRANSCRIPTION

+	Pauses
///	Pauses très longues
=	Liaison non standard remarquable
/…, …/	Hésitations entre plusieurs transcriptions
…-	Amorces de mot
*	Syllabe incompréhensible
***	Suite de syllabes incompréhensibles
###	Passage enregistré non transcrit
$$$	Coupure de l'enregistrement

ANNEXE 1 – ALPHABET SAMPA

CONSONNES	Symbole	Exemple	Transcription
Plosives	p	pont	po~
	b	bon	bo~
	t	temps	ta~
	d	dans	da~
	k	quand	ka~
	g	gant	ga~
Fricatives	f	femme	fam
	v	vent	va~
	s	sans	sa~
	z	zone	zon
	S	champ	Sa~
	Z	gens	Za~
Nasales	m	mont	mo~
	n	nom	no~
	J	oignon	oJo~
	N	camping	ka~piN
Liquides	l	long	lo~
	R	rond	Ro~
Semi-consonnes	w	coin	kwe~
	H	juin	ZHe~
	j	pierre	pjER

VOYELLES	Symbole	Exemple	Transcription
Orales	i	si	si
	e	ses	se
	E	seize	sEz
	a	patte	pat
	A	pâte	pAt
	O	comme	kOm
	o	gros	gRo
	u	doux	du
	y	du	dy
	2	deux	d2
	9	neuf	n9f
	@	justement	Zyst@ma~
Nasales	e~	vin	ve~
	a~	vent	va~
	o~	bon	bo~
	9~	brun	bR9~
	E/	= e ou E	
Indéterminées	A/	= a ou A	
	&/	= 2 ou 9	
	O/	= o ou O	
	U~/	= e~ ou 9~	