Laboratoire Ligérien de Linguistique

GUIDE DU TRANSCRIPTEUR

ESLOs

Version mai 2008

Table des matières

I. Objectifs du guide....4
II. Quelques mots sur les ESLOs....4
A. Logiciel de transcription :....5
B. Enregistrements des fichiers....5
IV. Principes de base :....6
V. Principes de segmentation....6
VI. Conventions de transcription....10
A. Enregistrement....10
1. Bruits....10
2. Durée de l'enregistrement....10
B. Principes d'écriture :....10
1. Signes graphiques....10
2. Trait d'union (segmentation lexicale)....12
C. Transcription....14
1. Orthographe....14
2. Épellation et sigles....14
3. Chiffres....15
4. Répétitions....15
5. Graphie incertaine....15
D. Interprétation....15
1. Passages peu compréhensibles....15
2. Ambigüités :....15
E. Etablissement des mots....16
1. Mots rétablis....16
2. Mots non rétablis....17
F. Les prononciations des mots étrangers (PP)....17
G. Soufflerie....18
1. Bruits de respiration....18
2. Clics....18
3. Point d'interrogation et ponctuation....18
H. Ponctuants....18
1. Enfin....18
2. Voilà....19
I. Affirmation et négation....19
1. Affirmation....19

2. Négation 19
J. Chevauchements 19
K. Onomatopées et de l'interjection (non-exhaustive) 21
L. Etape 1 22
M. Etape 2 22
N. Etape 3 22

PRÉSENTATION

I. Objectifs du guide

Dans ce guide vous trouverez l'ensemble des conventions qui vous permettront de transcrire les enregistrements de ESLO-1 et ESLO-2. Afin de vous guider au mieux dans cette tâche, une deuxième partie sera consacrée aux procédures de relectures et de vérifications de vos transcriptions. De plus, une version pilote (transcription de l'enregistrement 008) sera mise à votre disposition.

II. Quelques mots sur les ESLOs

ESLO-1, Enquête Socio-Linguistique à Orléans a été conduite à partir de 1968 par des universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Elle comprend environ 200 interviews, toutes référencées (caractérisation sociologique des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien), soit au total plus de 300 heures de parole incluant pour moitié des interviews en face à face et pour moitié une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.).

L'objectif est de mettre à disposition ce grand corpus oral en procédant à une première phase de transcription des données sonores. Actuellement, 120 heures ont déjà été transcrites.

En partant des acquis d'ESLO 1, une nouvelle enquête a été mise en chantier par le LLL : ESLO 2. Il s'agit à quarante années de distance de constituer un corpus comparable dans la produit attendu et dans les modalités de la collecte : l'objectif a été fixé à 400 heures environ de documents sonores répondant à une approche variationniste et conçus à travers des programmes spécifiques.

PHASE 1 : TRANSCRIPTION

III. Procédures de transcription

Seront mises à votre disposition chaque semaine 2 heures de transcription, le temps de transcription ayant été évalué à 15 fois le temps, soit 30 heures de transcription.

A. Logiciel de transcription :

L'outil choisi pour transcrire est **TRANSCRIBER** : logiciel d'aide à l'annotation de signaux de parole. Ce choix repose notamment sur son interface graphique simple permettant à un utilisateur non informaticien de segmenter des enregistrements de longue durée, de les transcrire et de marquer les tours de parole, la segmentation thématique et les conditions acoustiques, mais aussi pour sa robustesse face à de grand corpus. (Cf. doc : manuel d'utilisation du logiciel)

Pour chaque transcription, la procédure sera la suivante :

1. Avant de commencer vous aurez quelques paramètres à activer dans Transcriber :
   - Options générales : Nom du transcripteur, enregistrer l'activité dans (C:\time.txt)
   - Cette procédure permettra d'évaluer le temps de transcription pour les enregistrements ESLO.
2. Transcription en respectant les conventions
3. Relectures, vérifications

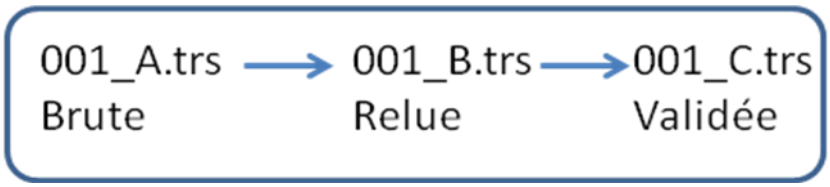
B. Enregistrements des fichiers

Les fichiers transcrits devront avoir la même dénomination que l'enregistrement sonore correspondant.

Pour certaines transcriptions, le travail donnera lieu à 3 versions :

1. A : transcription brute faite par un premier transcripteur.
2. B : relecture de la transcription brute.
3. C : validation de la transcription relue.

Nous obtiendrons ainsi :



Vous devrez également pour chaque transcription remplir les formulaires (remarques et problèmes) mis à votre disposition. Toutes ces informations alimenteront la base de données ESLO, ce qui permettra d'effectuer des requêtes sur ces données.

IV. Principes de base :

- Transcription orthographique : le dictionnaire sert de référence pour la graphie des mots.

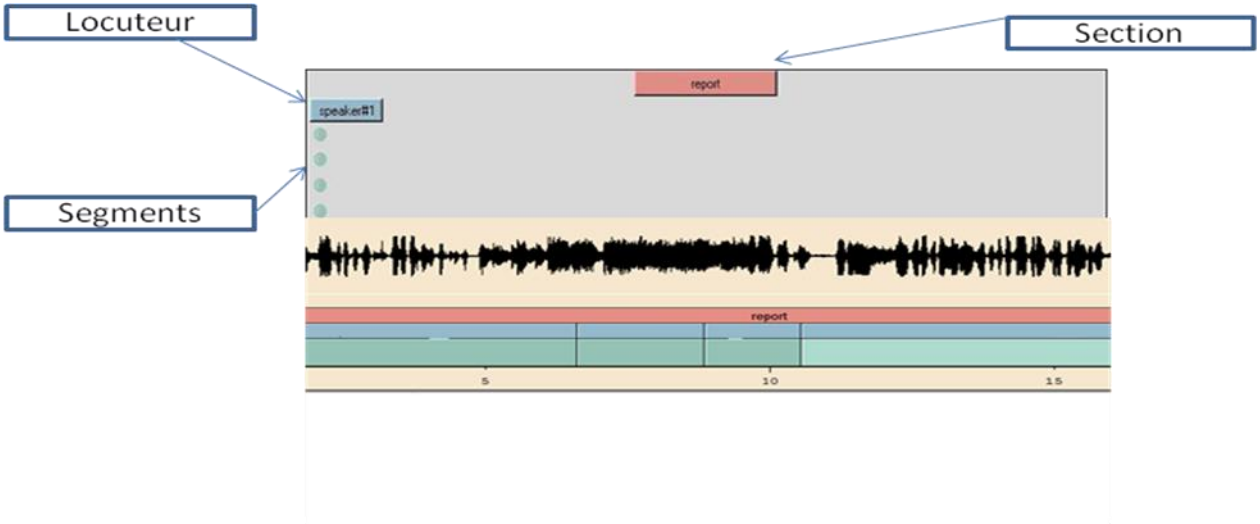
- **Règles de base à respecter :** *(certains de ces éléments seront repris de manière plus détaillée)*

- L'écriture des mots doit être homogène, aussi bien pour les noms propres que pour les noms communs.
- Ne pas utiliser la majuscule en début de phrase ni en début de tour sauf pour les noms propres.
- Excepté le point d'interrogation, aucun signe de ponctuation ne devra figurer dans vos transcriptions.

V. Principes de segmentation

Trois niveaux de segmentation sont possibles avec Transcriber :

- En sections
- En locuteurs
- En segment

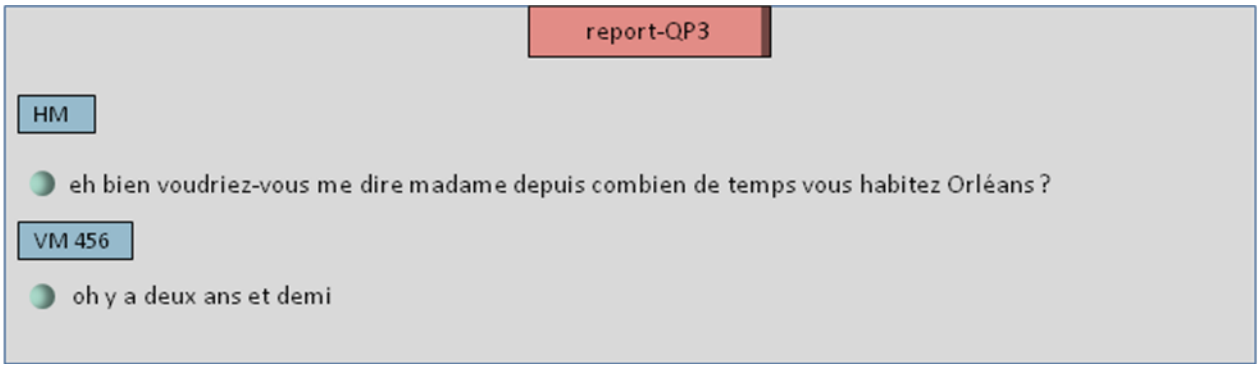


Quelques règles de segmentation seront à respecter :

1. Les sections :

Chaque section aura un code qui renverra à chaque question ou thème du questionnaire : QP1 (question préliminaire 1), S12 (section 1, question 2 :). Ainsi chaque report dans Transcriber englobera la question et la réponse.

Exemple :



2. Identification des locuteurs :

a. Locuteurs identifiés :

Pour l'identification des locuteurs il faudra se référer au document : « Catalogue des enregistrements des enquêtes sociolinguistiques à Orléans 1968 – 71 », pour ESLO-1 et une liste vous sera fournie pour ESLO-2.

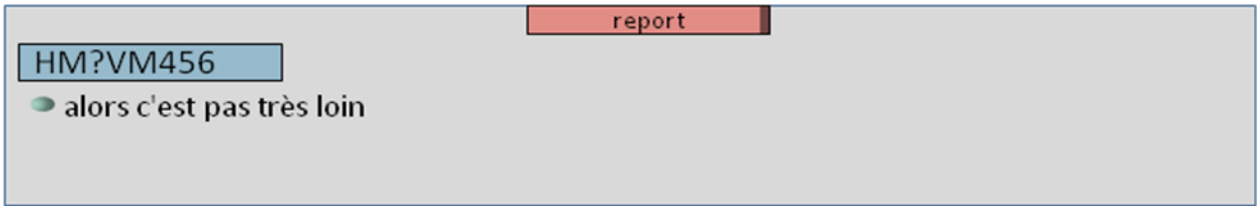
Ainsi, pour chaque enregistrement vous disposez, d'un code pour le témoin (ensemble de chiffres et de lettres) et d'un code pour le chercheur (initiales).

Prenons pour exemple l'enregistrement numéro 013 :

- Le témoin : MD 461
- Le chercheur : BV

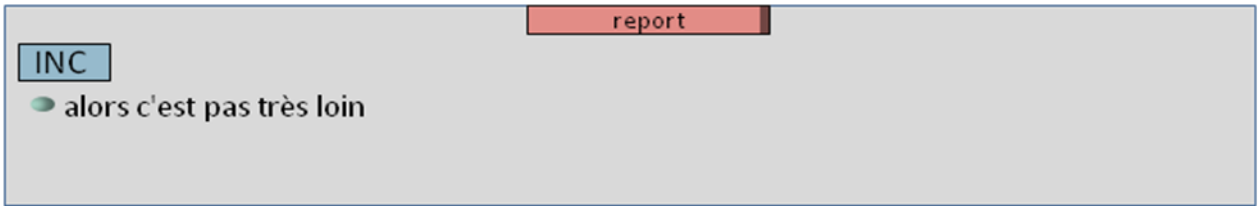
b. Hésitations entre deux locuteurs

Dans certains cas, il peut s'avérer difficile d'attribuer le tour de parole à l'un ou l'autre des locuteurs. De ce fait, cette hésitation sera notée par un point d'interrogation entre les deux locuteurs pour lesquels vous hésitez.



c. Locuteurs non identifiés

Lorsque le locuteur n'est pas identifié, il peut s'agir d'une personne extérieure à l'interview le transcripteur ne dispose ainsi d'aucunes informations le concernant. Dans ce cas, il sera noté par INC.

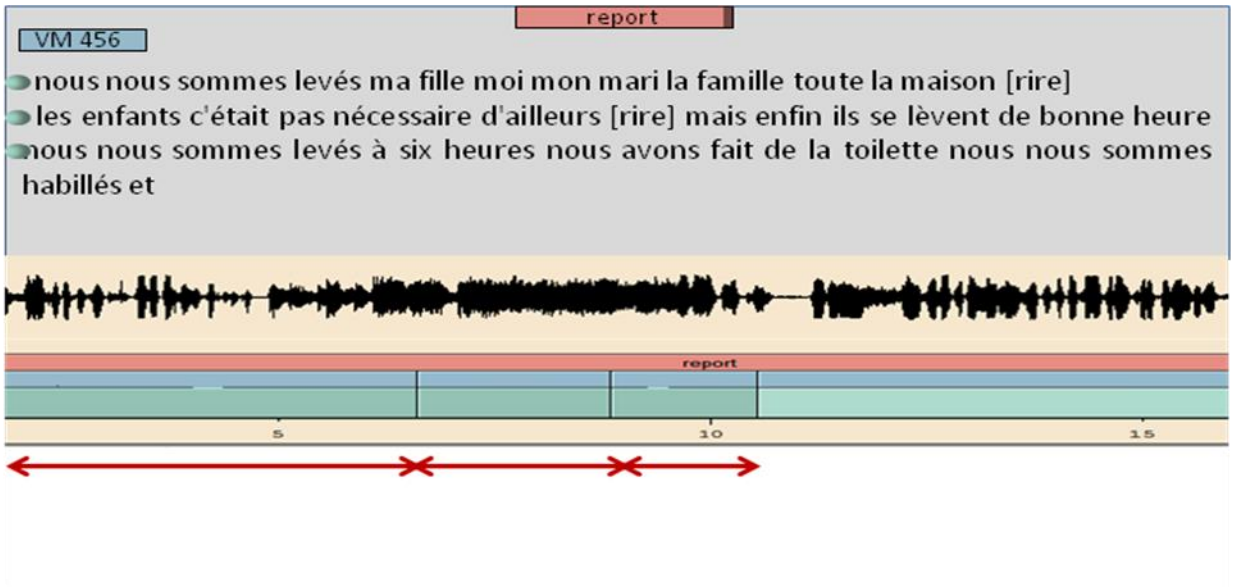


3. Segments :

La segmentation en unités doit se faire en respectant certaines règles :

- Les segments ne doivent pas être trop longs et ne doivent pas dépasser 15 secondes, comme c'est le cas dans l'exemple ci-dessous :

- Nous lui préférerons cette segmentation :



- Veillez à avoir des unités syntaxiques et sémantiques cohérentes
- Ne pas segmenter mot à mot mais par groupe pertinent.

Un des phénomènes de l'oral qui nécessite une segmentation particulière est la pause, elle est à noter par un segment vide. Cette segmentation permettra d'avoir précisément la durée de la pause.

VI. Conventions de transcription

Nous avons fait le choix d'une transcription semi-fine, qui permet de disposer d'une transcription rapidement en gardant certaines marques de l'oral.

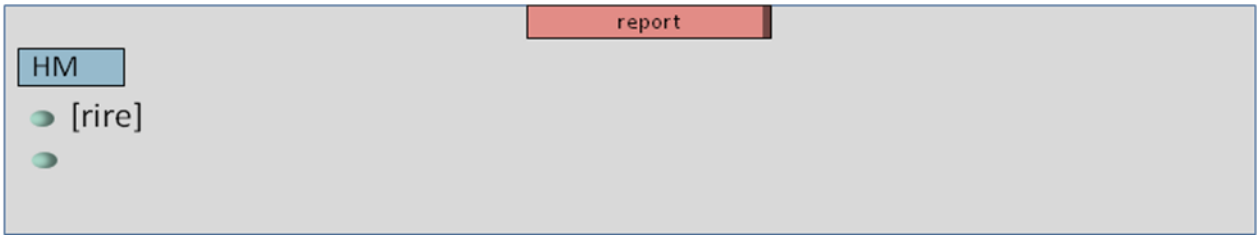
A. Enregistrement

1. Bruits

Les bruits sont à indiquer par les balises proposées par le logiciel, cependant ils ne seront pas tous systématiquement notés, nous maintiendrons essentiellement la balise [rire].

En revanche, il faudra adapter leur notation aux segments, si certains bruits ont une durée trop longue ou une incidence directe sur les réalisations des locuteurs.

Par exemple en (096/46:16) le bruit de micro est compris dans une séquence qui va de 46:12 à 46:20, alors que les quatre premières secondes sont une pause sans bruit de micro. Dans ce cas, il faudra isoler le bruit de micro dans un segment, en l'indiquant par une balise.



2. Durée de l'enregistrement

Dans le cas d'ESLO 1, les enregistrements de plus d'une heure sont coupés entre la soixantième et la soixante-cinquième minute. (109/01:04:03 > 1:04:24), (129/1:02:59), (133/1:03:50 > 133/1:04:01). Dans ce cas il faudra insérer un commentaire pour indiquer qu'il s'agit d'une coupure et non d'une pause : {*coupure enregistrement}*

B. Principes d'écriture :

1. Signes graphiques

a) Guillemets

Ne jamais en mettre, même dans les cas de discours rapporté.

On ne transcrit pas :

(…) veut dire par « on » ?

MAIS

(…) veut dire par on ? (107/19:45)[1]

b) Apostrophe

(1) Utilisation de l'apostrophe

Afin de préserver la reconnaissance automatique des unités, l'apostrophe ne doit être utilisée que lorsqu'elle peut correspondre à un usage orthographique.

- D'où y'a > y a

Car il n'y a pas de suite « y' » en graphie standard.

En revanche, on admet :

- qu'y a eu un changement (107/8:59)

(2) Absence d'élision

A l'inverse, pas d'élision lorsque le schwa a un contenu mélodique devant voyelle sans qu'il puisse être pour autant assimilé à « euh ».

- parce **que** on (107/7:49)
- parce **que** il (109/4:51)
- **que** ici (118/01:57)

(3) Apostrophe non suivie d'un mot

Un mot inachevé peut se terminer par l'apostrophe si telle était la forme qui eût été réalisée, l'apostrophe remplissant dans ce cas la même fonction que le tiret notant un mot inachevé.

de bien l' de bien l'écrire (139/18:24)

[1] Enregistrement 107, 19mn45

2. Trait d'union (segmentation lexicale)

a) Usage du trait d'union

Conventions exactes du français qui s'applique à tous les types de nombre :

- demi-heure (118/02:37)
- trois-quarts d'heure (118/02:38)
- mille neuf cent vingt-six (129/8:36)

Exceptions possibles :

- dictionnaire français anglais (107/32:04)
- rue du Maréchal-Foch > rue du Maréchal Foch (109/42:05)

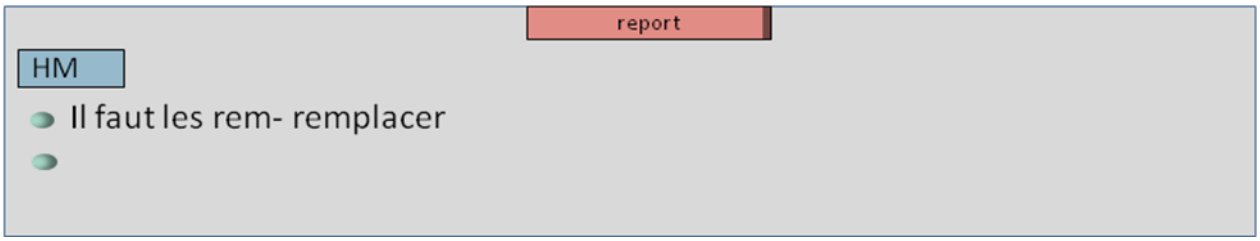
On distingue :

- cette place-là (106/12:13) de « cette place là »

b) Mots incomplets

Le trait d'union est la notation arrêtée pour les mots incomplets, cas ou le locuteur commence un mot et le termine pas. Vous ajouterez dans ce cas un tiret accolé à la partie tronquée, et veillez à laisser un espace entre le tiret et le mot qui suit.

Exemple :



c) Segmentation

En revanche, la segmentation au milieu d'un mot n'est pas admise.

- des Port- -ugais > des Portugais (129/39:44)
- sa- -voir-livre (129/1:20:19) > savoir-livre

d) Majuscules

(1) Emploi lexical

Elles ne sont pas utilisées pour marquer le début de phrase mais sont conservées pour les noms propres, comprenant les noms de personnes et de lieux, les noms d'institutions et de marques.

- Caisse Nationale d'Epargne (078/10:49)
- Waterman (106/47:30)
- Tuperware (107/31:09)
- Martini (118/32:59)

Problème des institutions qui sont spécifiques :

Deux cas particuliers et d'emploi fréquent (à cause du questionnaire) :

- sécurité sociale (106/53:28)

MAIS

- Académie Française (118/39:12)
- Bureau > Bureau (118/07:49) pour le BRGM
- Ponts et Chaussées (129/1:07:03)
- l'Ecole Centrale (131/10:31)
- bons du Trésor (133/1:01:06)
- une Université Nouvelle (542/39:54

(2) Assimilation à des noms propres

- « le Petit Larousse », en gardant la majuscule à l'adjectif (et donc, « le Grand » (106/35:17)
- « le Memento » (109/04:15).
- bic, parce qu'il est pris pour un équivalent générique de stylo bille opposé à stylo plume (096/40:23) (107/53:17) (118/41:35)

Sont aussi notés avec des majuscules :

- **les groupes de musique :** « Les Haricots Rouges » (096/55:33),
- **les titres de films :** « Le Cerveau » (096/56:52),
- **les titres de périodiques :** « les Sélection », i.e. *Sélection du Reader's Digest* (096/1:02:02), « la Nouvelle République » (109/48:03), « Tout l'Univers » – collection encyclopédique par fascicules – (118/34:32)
- **les modèles de véhicule ou d'avions :** Concorde (110/46:19) avec l'exception d'« une deux chevaux » (129/18:59)
- **les événements historiques :** la Guerre de Quatorze (110/51:20) (129/1:14:40), la Première Guerre Mondiale (129/47:49), la Libération (133/52:53)
- **les sites et monuments :** Gare du Nord, Palais Royal (110/52:09 > 52:14) mais, à Orléans, « le parc floral » (118/30:21) vs « Floralies » (118/24:12).
- **Les nationalités :** s'il ne s'agit pas de l'adjectif, prennent la majuscule : « Anglais » (107/30:56), en revanche pas de majuscule à « dieu » dans mon dieu » (118/00:16).

C. Transcription

1. Orthographe

Sauf s'il devait en résulter une contradiction entre ce qui a été dit et ce qu'on écrit, il faut respecter les conventions orthographiques :

- « des convictions que j'ai eues » (à propos de 106).

On écrit :

- et bien > eh bien (*passim*)
- oh la la > oh là là (096/40:09)
- où que c'est beau > ouh que c'est beau (109/12:55)

Attention aux difficultés de l'orthographe :

- un mot-croisé > un mots-croisés (132/44:59)

2. Épellation et sigles

Les lettres sont à noter en capitales.

Pour les épellations, elles seront notées avec des espaces entre les lettres.

- H A R I C O T

En revanche les sigles sont notés avec les lettres accolées et sans point :

- TVA > TVA (542/1:04:46)
- CGT > CGT (542/1:07:08)

3. Chiffres

Ils sont à noter en toutes lettres.

- une quatre cent quatre Peugeot (118/13:12, 14:19, 14:30)

Cependant lorsque les chiffres dépendent d'une suite de lettres, ils seront notés en chiffres.

- appartement F quatre > appartement F4

4. Répétitions

Tous les termes répétés sont transcrits.

- tout tout tout tout tout

5. Graphie incertaine

Lorsque vous êtes confrontés à un terme dont vous ne connaissez pas la graphie (nom propre, nom commun, onomatopées…) et que vos recherches ne vous ont pas permises de trouver la graphie exacte, vous devrez dans ce cas accoler au mot le caractère &.
Cette convention s'applique également à tous les mots non lexicalisés. (verlan, mot inventé…)

D. Interprétation

1. Passages peu compréhensibles

Ils sont à noter avec la balise prononciation inintelligible proposée par Transcriber.

(129/02:16, 35:38, 1:11:24) (215/04:58).
[pron = pif] beaucoup (106/20:06)

En revanche, il faut bien distinguer la balise [pron=pif] -> prononciation inintelligible faible, de la balise [pron=pi] -> prononciation inintelligible.

2. Ambigüités :

Lorsque vous hésitez entre 2 formes, le recours au contexte devrait vous permettre de faire certains choix (non exclusifs), si cela ne suffit pas choisissez la forme qui selon vous convient le mieux.

- ça changeait > ça a changé (106/32:41)
- ça la fatiguait > ça l'a fatiguée (109/16:00)

E. Etablissement des mots

1. Mots rétablis

a) Elisions erronées

Dans ce cas, vous devez rétablir la seule possibilité admise par l'orthographe.

- le petit qu'était malade > le petit qui était malade (106/14:01)
- y a tout ce qui faut > y a tout ce qu'il faut (109/28:51)
- qu'a son CAP > qui a son CAP (110/11:37)
- qu'était > qui était (118/01:49)
- tout ce qui fallait > tout ce qu'il fallait (139/14:23)
- t'as appris > tu as appris (110/37:38)
- t'écris > tu écris (110/44:59)
- ce qu'y reste > ce qu'il reste (109/52:15)

b) Suppression indécidable

Vous devez préserver la notation d'une unité si l'enregistrement ne permet pas de décider de la suppression effective de l'unité.

- qui y a déjà > qu'il y a déjà (096/1:07:27)
- tout ce qui y a > tout ce qu'il y a (139/07:15)
- guide savoir-vivre > guide de savoir-vivre (129/1:19:52)

c) Aphérèses et apocopes

Les aphérèses sont rétablies, quittes à faire l'objet d'une balise prononciation.
Les apocopes ne sont pas rétablies (elles apparaîtront dans le lexique juste avant l'unité concernée).

- tendez > attendez (129/3 :08)
- fin > enfin (passim)
- mais
- perfo (109/09:03) pour : *perforatrice*
- perfo-vérif (109/09:26) pour : *perforatrice-vérificatrice*

d) Rétablissement de la forme standard

- pu > plus (132/43:41)

2. Mots non rétablis

a) Lapsus

Lorsque la forme existe dans le lexique, le lapsus ne doit pas être corrigé, en revanche quand ce n'est pas le cas, vous devez la corriger et indiquer dans la balise [pron=], la forme prononcée par le témoin.

- omelette (109/00:37) *[pron=oblette]*
- faire la dastylo (106/05:25)
- préféreriez-vous (réf. ?)
- rémunéré > rénuméré (sic) (110/3:42) (129/24:22)
- menthélisées > mentholisées (129/30:13)

b) Déformations volontaires

Lorsque vous avez à faire à des formes délibérément déformées à des fins métalinguistiques, le procédé utilisé et le même que pour les lapsus.

- il fait beau [*pron = biau*] temps (110/23:53)

c) *Ne* de négation

Le premier terme de la négation ne doit être indiqué que s'il figure explicitement mais aussi si sa présence est indécidable du fait d'une liaison.

- on a pas > on n'a pas (*passim*)
- a fortiori, parce que le « y » a été prononcé /i/ et non /j/ :
- on n'y attache moins (08/41:07)

F. Les prononciations des mots étrangers (PP)

Les mots étrangers doivent être écrits de la même manière que dans leur langue d'emprunt.

- ketchup [pron= katchoup]

Certaines réalisations spéciales peuvent avoir lieu, dans ce cas vous utiliserez la balise prononciation pour signaler la prononciation réalisée.

G. Soufflerie

1. Bruits de respiration

Ils ne sont notés que s'ils donnent une indication sur l'expression d'un jugement.

- inspiration d'embarras de réponse,
- expiration d'agacement ou de désaccord

2. Clics

Les clics, quand ils sont bien audibles, peuvent être notés par une balise [bb] (bruit de bouche). Quelques exemples :

- (109/09:31, 39:04)
- (110/02:07)
- (133/46:42)
- (215/07:16, 09:06)

3. Point d'interrogation et ponctuation

Le point d'interrogation est la seule ponctuation maintenue car il est nécessaire pour désambiguïser :

On le porte systématiquement après :

- hein ?
- voyez-vous ? (118/02:07) (139/28:07, 30:54)
- comment dirais-je ? (118/43:53)

On porte aussi le point d'interrogation lorsque la question est suspendue au milieu en considérant que l'achèvement n'est pas nécessaire :

- bloc correspondance ou ? (106/56:5)

H. Ponctuants

1. Enfin

Conformément à ce qui a été dit concernant l'aphérèse, vous noterez :

- fin > enfin (110/2:45, 9:52) (133/08:06)

2. Voilà

Pas de forme réduite car elle ne semble pas distinctive par rapport à la forme pleine :

- Voilà > v'là (1:01:40)

I. Affirmation et négation

1. Affirmation

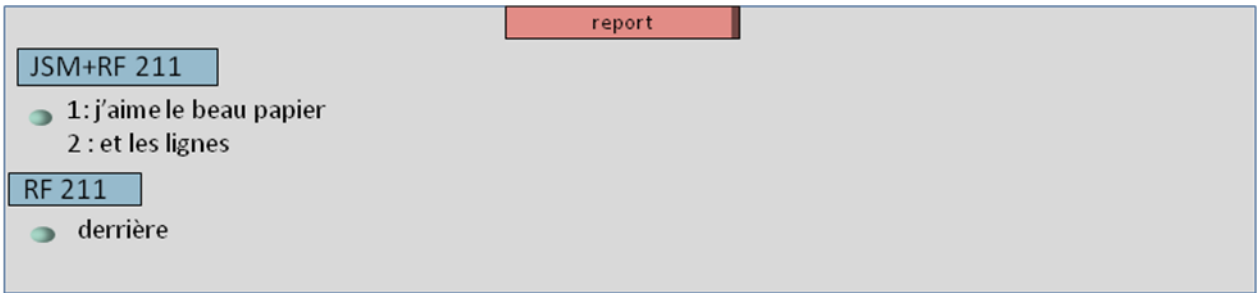
Les formes d'approbation sont actuellement partagées entre « oui », « ouais » et « mh » (ou plutôt « mh mh). On note « ouais » quand la substitution de « oui » modifierait la valeur du terme :
(106/49:21, 1:00:33)
(107/12:19)
(110/36:17)
(132/48:01)

2. Négation

Non sera toujours noté « non », et non « nan » (078/06:59, *passim*). Le « nan sera réservé aux formes enfantines du « non ».

J. Chevauchements

Lorsque 2 locuteurs parlent en même temps, vous utiliserez la segmentation que propose Transcriber (paramètres du tour>parole superposée). Il faudra seulement mettre la portion de parole prononcée simultanément et non l'ensemble de l'intervention des locuteurs.


report
JSM+RF 211
1: j'aime le beau papier
2 : et les lignes
RF 211
derrière

K. Onomatopées et de l'interjection (non-exhaustive)

Voici la liste des principales onomatopées et interjections que vous serez amené à rencontrer dans vos transcriptions :

ah	Cha	mah	poh
oh	chpoum	maille	pop
aha	chtac	miam	pouf
atchoum	clac	mouais	pouh
aïe	clou	moui	poum
baf	ding	of	snif
bah	Euh	ouah	teu
bam	gla	ouais	vouf
bap	gna	ouf	vouh
ben	gnin	ouille	voum
beuh	hein	oïe	vroum
bim	hop	pam	wahou
bla	hou	pff	ya
bloum	hu	pim	yeah
boah	hum	piouc	yoh
bof	hé	plaf	zut
boh	la	ploum	
bouh	leu	pof	
boum	lin		
bé			

PHASE 2 : RELECTURE

Une fois votre transcription terminée, plusieurs relectures sont nécessaires. Dans cette partie, vous trouverez un ensemble d'étapes qui vous permettront d'appréhender au mieux cette phase.

3 étapes sont nécessaires, chacune permet en effet de se concentrer sur un élément précis de la transcription.

L. Etape 1

La première étape consiste à exporter votre transcription en format texte à partir de Trancriber. Vous récupérez ensuite le fichier texte que vous copierez sur Word. Cette phase permet dans un premier temps de repérer les fautes de frappe et les fautes d'accord, mais attention ne vous fier pas uniquement à son correcteur, tous les éléments ne seront pas toujours repérés.

M. Etape 2

La deuxième étape consiste en une relecture sans le fichier sonore. Durant cette phase, il s'agira de vérifier les fautes d'orthographe, de grammaire, de conjugaison, d'accord, l'écriture des noms de ville, de pays…Vous devez veiller également à l'homogénéité de vos transcriptions, l'écriture des mots, des noms de ville, des néologismes doit être la même dans toute la transcription.

Pour vous aider dans cette phase, nous mettons à votre disposition une liste de sites à consulter :

- http://atilf.atilf.fr/tlf.htm : « Le trésor de la langue française informatisé », qui comprend un correcteur d'erreurs, une liste de mots, et enfin une saisie phonétique qui permet de retrouver un mot dont on a que la prononciation.
- http://www.linternaute.com/dictionnaire/noms-propres/recherche/ : permet de rechercher des noms de villes, de pays.
- http://mapage.noos.fr/mp2/ : Bréviaire d'Orthographe Française
- http://www.leconjugueur.com/ : conjugaison du verbe que l'on souhaite, ainsi que règles de grammaire.

N. Etape 3

Cette dernière étape consiste à une vérification avec le fichier sonore, cette phase vous permettra de porter votre attention sur la segmentation, et le respect des conventions (amorces, répétitions…).

Version mai 2008