### Version définitive des rubriques des métadonnées remplies de DB04

=====================================================================

Sur les fichiers eux-mêmes (fichier audio fusionné .mp3 et le fichier .cha qui contient la transcription alignée avec l’audio)

Le fichier audio mp3 déposé est une fusion à partir des fichiers segmentés fournis par le RAHMI (car l’alignement n’était pas possible à partir du fichier mp3 unique fourni par le RAHMI).

Il reste 3 balises <audiomanquant> qui correspondent à des passages manquants dans les fichiers segmentés et donc fusionnés.

L’alignement entre les fichiers texte et audio a été effectué avec un séquençage de 20 secondes environ.

Informations générales sur le mode de transcription

Afin d’évaluer les variations entre transcripteurs, une écoute par un 2ème transcripteur a été effectuée sur la version transcrite et annotée telle que déposée. Les variations sur les premières lignes du fichier texte (environ 100 lignes) (version déposée → remarque ou autre proposition d’un 2ème transcripteur) (les numéros de lignes sont indicatifs et susceptibles d’avoir changé en fonction de corrections ultérieures) :

* l.18 “<pause>” → “<pause> <mic>”
* l.19 “<revision> était donc euh </revision>” → “<revision> était donc <euh> </revision>” car balises englobantes possibles (cf. mail sur le sujet : pris en compte dans le guide transcription ?)
* l.22 “<emphase> oui oui oui </emphase> oui” → “<emphase> oui oui oui </emphase>”
* l. 23 “il a été à l'école militaire” → bafouille sur “été” mais pas de balise prévu pour cela
* l.36 “</repetition> <silenceentendu>” → “</repetition> <bb> <silenceentendu>”
* l. 37 “<mdd> mais alors là </mdd> j'en suis pas sûre” → “mais <mdd> alors là </mdd> j'en suis pas sûre”
* l. 39 “<revision> de de </revision>” → “<mic> <revision> de de </revision>”
* l.42 “peut-être <euh>” → <euh> entre le euh et le pff
* l.42 “six mois je ne sais pas” → “tss” ou “c” avant le “je”
* l.50 “déjà et” → “déjà <bb> et”
* l.59 “ils étaient deux ou trois <revision> je peux pas vous dire que </revision>” : est-ce que c’est une révision ou un mdd ?
* l.61 “Marseille” → prononciation un peu coupée “Marsei-”
* l.65 “d'argent et” → “d'argent <bc> et” (mais indiqué dans les métadonnées que tous les <bc>ne sont pas notés car nombreux)
* l.68 “<mdd> oh ben </mdd> va de l'autre côté” → “<bruit> <mdd> oh ben </mdd> va de l'autre côté” car sonnerie de téléphone pendant qu’elle parle. Voir aussi si <bruit> ou utiliser plutôt <bruitext> pour éviter la reconnaissance automatique + “va de l’autre côté” ne concerne pas le récit en <mdd> ?
* l.69 “<mdd> alors </mdd>” → “<mdd> alors i- </mdd>”
* l.72 “le bûcheron qu'il avait jamais” → j’entends aussi “le bûcheron qu'il n’a jamais” mais pas sûre
* l. 76 “<revision> avec euh </revision>” → “<revision> avec <euh> </revision>” à voir ce qui correspond au guide
* L. 84 “<revision> 40 euh 30 ne 41 42 </revision>” → “39 <pron=trentene” car certainement cette année-ci. Quel formalisme “30-” ? (même si ici dans révision)
* l.91 “je crois” → en <mdd> ou <revision> ?
* l. 92 des <bc> présents
* l. 104 “c'est 42 là je” → “c'est <css> 42 là je” <css> ou “tss”
* l.105 “<revision> f- par le </revision>” → “<revision> <pff> par le </revision>” (pff ou fff)
* l. 107 “je peux pas vous bien vous” → ou selon interprétation : “je peux pas <revision> vous </revision> bien vous”
* l. 109 “il avait des contacts” → “<repetition> il </repetition> il avait des contacts”
* l. 113 “résistance ils étaient commandés” un <bc> après “résistance”
* l. 117 “y avait déjà des groupes → “<repetition> oui </repetition> oui <bc> y avait déjà des groupes”
* l. 128 “on est partis à ” → “on est partis <repetition> à </repetition> à”

Contexte de l’enregistrement

L’enregistrement audio n’a pas était le plus évident à transcrire, annoter et aligner dû à de nombreux chevauchements de parole. Des confusions entre les interlocuteurs sont possibles, par ex. DB04FEM indiquée à la place de CH\_MB1.

Plus que pour les autres récits, la transcription déposée est différente des premières versions (dont celle du RAHMI) : beaucoup de mots en trop ou alors à l'inverse non transcrits, des passages non ou mal annotés, des fautes d'orthographe et de conjugaison.

Toutes les balises <bc> ("BackChannel") ne sont pas notées car trop nombreuses.

Remarques sur l’oral (en relation avec le guide de transcription)

A priori, beaucoup de révisions et d’interjections. Pour certaines révisions, elles sont annotés sur le 2ème segments (et non le 1er comme habituellement), par ex *décembre 45 <revision> non dé- décemb- oui </revision>* ; *quand mon mari <revision> a vu ah oui alors quand il </revision> était à Marseille*.

Quelques prononciations systématiques :

* DB04FIE prononce systématiquement "guérilleros" à l'espagnol, donc ce n’est pas indiqué par une balise <pron=Hisp>.

AJOUTER REMARQUES CARMEN

Incertitudes dans la transcription mais une interprétation est choisie

Interprétation des Npr de lieux et de personnes

Pour les noms de lieux :

Des choix de transcription même si incertitude :

* *Marciac <pron=marsac>* au vu de la chronologie des évènements et du fait que Marciac est cité à d’autres moments dans le récit. Toutefois, Marciac qui est dans le Gers et Marsac qui est dans les Hautes-Pyrénées sont situés à une vingtaine de kilomètres l’un de l’autre.

Pour les noms de personnes :

A VOIR :

* 292-301 : NP153 et NP154

<DL> Il semble que ce soit une seule et même famille. Donc un seul et même nom de famille (nom composé) ?

<LJ> c’est vrai que c’est exprimé comme une même famille, mais c’est peut-être deux familles liées par un mariage et du coup ce sont quand même deux noms séparés ?

### Commentaires Déborah puis Carmen, Catherine, Laurence

=====================================================================

Commentaires Déborah/réponses, et Carmen, Catherine, Laurence

Notes Déborah

===========

DB04

Fichier audio

=========

Fusion à partir des fichiers segmentés (car alignement impossible à partir du fichier mp3 unique fourni par le Rahmi)

Remarques Deborah

===============

Ajout du header (@Begin ... @End)

Ajout d'une \* à chaque tier

Alignement avec séquençage de 20 secondes environ

Balisage des NprLieu avec les identifiants de gazetier : ok

Vérification des codes de noms propres de personnes : ok

==================================================================================================================

Corrections systématiques :

==================================================================================================================

Tous les [] remplacés par des <> (<euh>, <silence>, <pause>, <pron=nonstd>, <pron=pi>...)

<mdd> euh/hein </mdd> --- <euh/hein>

<mdd> et bien </mdd> --- <mdd> eh bien </mdd>

<mdd> mais/oui </mdd> ? >> corriger si nécessaire

<mdd> ah/eh/hé/oh/bah </mdd> --- <ah/eh/hé/oh/bah>

<rire/silence/pleur> --- <rireentendu/si/pleurentendu>

<revis> </revis> --- <revision> </revision>

<espagnol></espagnol> --- <lg=espagnol></lg=espagnol>

XXX là --- XXX-là

<audiomanquantsep/fus> --- effacés

Ajouts de majuscules, par exemple :

les républicains --- les Républicains

les espagnols --- les Espagnols

==================================================================================================================

Corrections notables :

==================================================================================================================

<DL> 159 : la <lg=espagnol> Retirada </lg=espagnol> <trad> </trad>

J'ai annoté ainsi, mais je ne suis pas sûre, étant donné qu'on ne le traduit pas en français par "Retraite".

<LJ> plutôt ne noter aucune balise comme Retirada est dans le lexique MATRICIEL

<DL> 559-571 : juste un "gros" exemple de quand on groupe tout dans <révision>.

<LJ> Il y a plusieurs balises <revision></revision>, mais cela me semble bien...

==================================================================================================================

Questions/Réponses :

==================================================================================================================

<DL> 426-428, 473-474, 629 : Il reste juste trois <audiomanquant>, donc bien moins qu'avant, qui correspondent à des passages manquants dans les fichiers segmentés.

<LJ> ok

292-301 : NP153 et NP154

<DL> Il semble que ce soit une seule et même famille. Donc un seul et même nom de famille (nom composé) ?

<LJ> c’est vrai que c’est exprimé comme une même famille, mais c’est peut-être deux familles liées par un mariage et du coup ce sont quand même deux noms séparés ?

320, 326, 333, 336... : DB04FIE prononce systématiquement "guérilleros" à l'espagnol

<DL> J'indique <pron=Hisp> ? Ou dans les méta-données ?

<LJ> je dirais dans les métadonnées si c’est systématique.

441-442 : décembre 45 <revision> non dé- décemb- oui </revision>

542-543 : quand mon mari <revision> a vu ah oui alors quand il </revision> était à Marseille

<DL> J'ai placé <revision> sur les deuxièmes segments et non les premiers.

<LJ> ok pour le 1er exemple, le 2ème exemple est sur le premier segment ?

391 "Marsac"

<DL> Je dirais qu'elle parle de Marciac (vu la chronologie des évènements, et c'est aussi la seule occurrence de Marsac dans tout l'enregistrement), mais qu'elle ne le prononce pas correctement. J'écrirais donc :

Marciac <pron=marsac>

Gazetier :

<DL> Il est écrit "Ébre" dans le gazetier, mais est-ce que ce n'est pas plutôt "Èbre" (avec un accent grave) ?

<DL> Il n'y a pas d'occurrence de "Ebro" dans l'enregistrement et la transcription (il y a seulement "Ébre").

<DL> Si on est d'accord pour Marsac/Marciac, il n'y a plus besoin de l'id de Marsac.

<LJ> Oui d’accord pour Marciac, c’est le plus logique, peut-être indiquer une balise sur une prononciation non standard ou avec la prononciation phonétique (note : Marciac dans le Gers et Marsac dans les Hautes-Pyrénées sont situés à une vingtaine de kilomètres l’un de l’autre).

Pour “Ébre” c’est une erreur, il s’agit bien de Èbre, et je supprime “Ebro” du gazetier.

Je modifie le gazetier en conséquence et remplace l’ancienne version sur le FTP.

<DL> Observation générale : Ce n'est pas l'enregistrement le plus évident. Il y a beaucoup de chevauchements de parole donc la transcription, l'annotation et l'alignement étaient moins évidents à effectuer.

Et ce n'était pas la transcription la plus évidente non plus : confusion des interlocuteurs (DB04FEM indiquée à la place de CH\_MB1...), et, plus que d'habitude : des mots en trop ou à l'inverse non transcrits, passages non ou mal annotés, fautes d'orthographe et de conjugaison...

<LJ> Je pense que l’on pourra prévoir de joindre ces informations aux métadonnées des fichiers pour le dépôt futur.

==================================================================================================================

Remarques à voir ensemble à la fin ou pour la suite :

==================================================================================================================

<LJ> 21/10/2019 : écoute DB04.cha (656 lignes) ligne 1 à ligne 128, version de Déborah du 18/10/2019

Remplacer les <pause> et <silenceentendu> par <si>

l.18 “<pause>” → “<pause> <mic>”

l.19 “<revision> était donc euh </revision>” → “<revision> était donc <euh> </revision>” (suite au mail du 21/10 sur balises englobantes)

l.22 “<emphase> oui oui oui </emphase> oui” → “<emphase> oui oui oui </emphase>”

l. 23 “il a été à l'école militaire” : bafouille sur “été” (mais pas de balise prévu pour cela)

l.36 “</repetition> <silenceentendu>” → “</repetition> <bb> <silenceentendu>”

l. 37 “<mdd> mais alors là </mdd> j'en suis pas sûre” → “mais <mdd> alors là </mdd> j'en suis pas sûre”

l. 39 “<revision> de de </revision>” → “<mic> <revision> de de </revision>”

l.42 “peut-être <euh>” : entre le euh et le pff

l.42 “six mois je ne sais pas” : “tss” ou “c” avant le “je”

l.50 “déjà et” → “déjà <bb> et”

l.59 “ils étaient deux ou trois <revision> je peux pas vous dire que </revision>” : est-ce que c’est une révision ou un mdd ?

l.61 “Marseille” : prononciation un peu coupée “Marsei-”

l.65 “d'argent et” → “d'argent <bc> et” : indiqué dans les métadonnées : tous les <bc>ne sont pas notés car nombreux.

l.68 “<mdd> oh ben </mdd> va de l'autre côté” → “<bruit> <mdd> oh ben </mdd> va de l'autre côté” : sonnerie de téléphone pendant qu’elle parle, voir aussi si <bruit> ou utiliser plutôt <bruitext> pour éviter la reconnaissance automatique + “va de l’autre côté” ne concerne pas le récit en <mdd> ?

l.69 “<mdd> alors </mdd>” → “<mdd> alors i- </mdd>”

l.72 “le bûcheron qu'il avait jamais” : j’entends aussi “le bûcheron qu'il n’a jamais” mais pas sûre

l. 76 “<revision> avec euh </revision>” → “<revision> avec <euh> </revision>” : à voir ensuite pour tout baliser du coup (mail du 21/10)

L. 84 “<revision> 40 euh 30 ne 41 42 </revision>” : “30 ne” pour “39” certainement, quel formalisme “30-” ? (même si ici dans révision)

l.91 “je crois” : en <mdd> ou <revision> ?

l. 92 des <bc> présents

l. 104 “c'est 42 là je” : “c'est <css> 42 là je” <css> ou “tss”

l.105 “<revision> f- par le </revision>” → “<revision> <pff> par le </revision>” (pff ou fff)

l. 107 “je peux pas vous bien vous” ou selon interprétation : “je peux pas <revision> vous </revision> bien vous”

l. 109 “il avait des contacts” → “<repetition> il </repetition> il avait des contacts”

l. 113 “résistance ils étaient commandés” un <bc> après “résistance”

l. 117 “y avait déjà des groupes → “<repetition> oui </repetition> oui <bc> y avait déjà des groupes”

l. 128 “on est partis à ” → “on est partis <repetition> à </repetition> à”