# NOUVEAU FORMAT DES RENDUS

**Version 1 : Jacques Durand, Bernard Laks et Chantal Lyche**

**Version 2 : Julien Eychenne, Philippe Hambye et Géraldine-M. Mallet**

(Dernière révision : janvier 2004)

Le projet PFC est basé sur un ensemble d'enquêtes réalisées de façon décentralisée dans le monde francophone. Ces enquêtes doivent être assez cohérentes et substantielles pour permettre les recherches présentes et à venir.
En termes scientifiques, nous nous sommes assignés deux phases principales :

1. 2002-2005 : La première phase vise à constituer une base de données d'environ 400 locuteurs et à proposer des premières exploitations des enquêtes. Dans la mesure où PFC est un projet à long terme, la base de données sera conçue de manière à pouvoir accueillir des nouveaux résultats d'enquête, au-delà des échéances fixées pour une exploitation de grande envergure.
2. à partir de 2005 : la deuxième phase consistera à exploiter les données à grande échelle. Pour ce faire, la base de données sera rendue homogène ; il est donc essentiel que les données recueillies dans différents points d'enquête répondent à une série d'exigences strictes quant à leur format.

Les rendus qui seront ainsi transférés aux centres de coordination (ERSS et MODYCO) par les équipes prenant part au projet contiendront donc l'ensemble des éléments mentionnés ci-dessous et uniquement ceux-là. Nous n'intégrerons à la base de données que les enquêtes qui sont *numérisées*, *transcrites* et *codées* pour les phénomènes communs au projet (à savoir le schwa et la liaison).

Rappelons que, pour chaque témoin, nous envisageons environ 1h d'enregistrement en tout, correspondant à la lecture de la liste de mots et du texte, à la conversation guidée (15-25 min. environ) et à la conversation libre (20-30 min.).

Nous attirons l'attention des participants au projet sur le fait que l'ancien format des rendus exigeait une fiche d'enquête pour chaque locuteur, et une fiche « note d'enquête » pour chaque point d'enquête. Puisque ces informations doivent être directement saisies en ligne sur le site du projet (http://infolang.u-paris10.fr/pfc), via des formulaires HTML, ces fiches sont dorénavant obsolètes et ne devront par conséquent pas être intégrées lors du rendu final des enquêtes. Nous suggérons que les enquêteurs complètent les formulaires pour les différents locuteurs dès les enquêtes terminées, ceci afin de prendre conscience des informations lacunaires à un moment où il est encore aisé de les recueillir.

## A. Rendu des données sonores

Pour chaque locuteur, il nous faudra :
4 fichiers .wav correspondant à la liste de mots, au texte, à la conversation guidée et à la conversation libre.

Les paramètres d'enregistrement sont les suivants :

Canaux : Mono
Taux d'échantillonnage : 22 Khz (22050 Hz)
Taille d'échantillonnage : 16 bits

Pour identifier les fichiers, nous avons mis au point un étiquetage où chaque locuteur et chaque activité sont identifiés de façon unique par une séquence de 8 symboles alphanumériques. Nous partirons de trois exemples hypothétiques avant d'examiner les principes sous-jacents à ces noms :

**31cmd1mw.wav** : fichier son d'un enregistrement fait dans le département français 31 (Haute Garonne), au point d'enquête c (Toulouse banlieue), du témoin md1 (Marie + Delomb + 1) qui lit la liste de mots (m) sous sa forme sonore (w), information qui sera également codée par l'extension .wav.
**cqajp1gw.wav** : fichier concernant une enquête faite au Canada (c) dans le Québec (q), point d'enquête a (Montréal), témoin jp1 (Jean-Luc Palerme 1) contenant la conversation guidée (g)
sous forme sonore (w).
**sgabg3lw.wav** : fichier son d'un enregistrement effectué en Suisse (s), à Genève (g), au point a, témoin bg3, Blanche Giraud indice 3 car l'enquête inclut aussi Bernard Giraud (bg1) et Bernadette Giraud (bg2); le fichier concerne la conversation libre (l) sous sa forme w, ce qui est également codé par l'extension .wav.

Les principes sont les suivants :

**Position 1 et 2** : pour la France, le département (01, 31, etc.), pour les autres pays l'initiale *b* (Belgique), *c* (Canada), *s* (Suisse), etc., suivie d'une initiale pour la ville ou la région.
**Position 3** : l'indice du point d'enquête, à savoir la lettre *a* s'il n'y a qu'une enquête, sinon *b* à *z* pour les divers points d'enquête par département ou pays (après accord entre les enquêteurs et la coordination du projet pour éviter des homonymies).
**Position 4, 5 et 6** : initiales du témoin prénom + nom + chiffre. Le chiffre est 1 s'il n'y a qu'un seul témoin avec les initiales en question. Au-delà de 1, les chiffres sont assignés en fonction de l'ordonnancement alphabétique des prénoms.
**Position 7** : m (pour les mots de la liste), t (pour le texte), g (pour la conversation guidée), l (pour la conversation libre).
**Position 8** : w (pour wave), g (pour TextGrid, voir ci-dessous). A ce niveau, notre codage est en partie redondant mais fournit des noms de fichiers plus transparents sans avoir à en examiner l'extension.

NB : Lors de la numérisation de la liste de mots et du texte, il est impératif de supprimer les commentaires qui précèdent, interrompent et suivent la liste ou le texte à proprement parler. D'autre part, nous n'intégrerons pas à la base de données commune les listes ou textes complémentaires spécifiques à certaines enquêtes.

## B. Rendus des transcriptions

Pour chaque locuteur, il faudra fournir 4 fichiers TextGrid, autrement dit un fichier par tâche effectuée par le locuteur.

Nous demandons des transcriptions orthographiques alignées sous PRAAT de 5 minutes pour la conversation guidée et de 5 minutes pour la conversation libre. Attention, ces fichiers, ainsi que les fichiers son associés, ne doivent pas être découpés en sous-fichiers. La transcription orthographique se fera à l'intérieur de la première tire qui apparaît lors de la création du fichier TextGrid associé au fichier son. Rappelons que cette tire doit être nommée : le label aura par exemple la forme « 31cmd1_transcript-graphe » (cf. *Bulletin PFC n°1*, 2002).

On notera qu'il faut également transcrire le texte lu sous Praat. En effet, le codage du schwa et de la liaison se fait à partir de l'alignement texte lu/son sous Praat pour trois raisons :

a) les lectures donnent lieu à des répétitions, des omissions ou des écarts qui sont importants pour les codages,
b) un codage réalisé directement dans des fichiers .txt ou .doc ne permet pas une intégration et des révisions dans Praat,
c) les outils développés au sein du projet ne fonctionnent qu'avec des fichiers au format TextGrid.

Enfin, nous demandons également un alignement texte/son de la liste de mots, afin de permettre des recherches automatiques sur les segments sonores correspondant aux différents items de la liste. La procédure à suivre consiste :

- à créer un fichier TextGrid associé au fichier son de la liste de mots (p. ex. 31adb1mw.wav),
- à placer une frontière (*boundary*) dans le TextGrid, au début et à la fin de chaque item de la liste (nombre + mot),
- à transcrire pour chaque item de la liste, le nombre en chiffres arabes et le mot dans l'intervalle correspondant. Les différents éléments d'un item seront séparés par un espace. On aura ainsi : 1 roc ; 6 fou à lier ; etc. En aucun cas, on n'aura recours à un point ( . ) ou un caractère de soulignement ( _ ). On veillera par ailleurs à ce que le nombre d'intervalles coïncide **exactement** avec le nombre d'items prononcés (à savoir 94), en "épurant" le fichier si nécessaire comme il a été rappelé plus haut. De même, on n'aura pas recours aux bornes DEBUT et FIN pour marquer le début et la fin de la liste de mots. Dans la mesure où les fichiers sonores doivent être "épurés", ces bornes sont superfétatoires.

Le rendu sera donc sous la forme de quatre fichiers du type TextGrid (sous Praat) signalés par la lettre *g* en huitième position des noms. Soit, par exemple, pour le témoin Marie Delomb (31cmd1) considérée plus haut :

31cmd1gg.TextGrid = transcription alignée de la conversation guidée (g) sous forme textgrid (g) avec extension .TextGrid
31cmd1lg.TextGrid = transcription alignée de la conversation libre (l) sous forme textgrid (g) avec extension .TextGrid
31cmd1tg.TextGrid = transcription alignée du texte lu (t) sous forme textgrid (g) avec extension .TextGrid
31cmd1mg.TextGrid = transcription alignée de la liste de mots (m) sous forme textgrid (g) avec extension .TextGrid

NB : Le choix de la localisation des 5 minutes à transcrire orthographiquement pour chacune des conversations est laissé aux responsables des enquêtes. Toutes choses égales par ailleurs, on commencera la transcription au début de l'enregistrement.

## C. Rendus analyses

Il y a deux types d'analyses communes à toutes les enquêtes : la liaison et le schwa. Attention, les codages de la liaison et du schwa se font sous PRAAT.

1) L'analyse du schwa porte sur 3 enregistrements : le texte, la conversation guidée (portion de 3 MINUTES) et la conversation libre (portion de 3 MINUTES). Dans une première phase, la création de nouveaux fichiers avait été envisagée. Il est apparu plus commode que les codages se fassent dans le ficher TextGrid principal sur des tires différentes. On créera donc une nouvelle tire en deuxième position dont le nom aura la forme « 31cmd1_schwa » (v. fonction « Duplicate tier » de Praat).

2) L'analyse de la liaison porte sur 3 enregistrements : le texte, la conversation guidée (5 MINUTES) et la conversation libre (5 MINUTES). Le codage de la liaison se fera sur une troisième tire créée comme précédemment à l'aide de la fonction « Duplicate tier » et nommé sur le modèle « 31cmd1_liaison ».

## D. Remarques importantes

- Attention les codages pour schwa et liaison se font sous Praat à partir de la transcription orthographique alignée. Gardez l'intégrité des fichiers TextGrid dans la mesure où les codages minimaux de départ seront sans doute élargis par la suite.
- L'analyse de l'inventaire phonologique du système de chaque locuteur qui avait été initialement prévue (à partir de la fiche établie à la section 2.2 des « Directions d'analyse ») est désormais abandonnée au profit de procédures automatisées. En effet, des outils développés au Laboratoire Parole et Langage (LPL, Université d'Aix-en-Provence) par Noël Nguyen permettront bientôt une analyse acoustique semi-automatique de la liste de mots, sur la base de l'alignement texte/son qui sera intégré dans les rendus.
- Les noms des enquêtés ne doivent pas apparaître dans les transcriptions. Ainsi, en considérant un locuteur fictif Pierre Hambert, on remplacera son nom de famille par l'initiale, de sorte que « Pierre Hambert » apparaisse dans le TextGrid comme « Pierre H. ».