FICHIERS MOTS : CONSTITUTION, ALIGNEMENT ET TRANSCRIPTION[1]

Jean-Michel Tarrier et Cyril Auran[2]
ERSS UMR 5610 CNRS & Université de Toulouse-Le Mirail

0. Introduction

L'objectif de cette courte notice est de présenter le format de rendu concernant la transcription des fichiers mots et l'alignement de celle-ci au signal. En effet, seuls les fichiers relatifs aux conversations et à la lecture de texte sont jusqu'à présent l'objet d'une " obligation " de transcription et d'alignement dans la version précédente du protocole. Depuis décembre 2002, il est convenu qu'il est également nécessaire d'avoir des alignements son/texte pour les listes de mots, sur lesquels reposeront les outils en cours de développement. Cependant, les éléments d'information relatifs au rendu de ces fichiers mots sont trop éparses et trop souvent méconnus. Le lecteur aura donc compris qu'il s'agit ici de remédier à cet état de fait et de rassembler les précisions indispensables pour un rendu plus achevé de ces fichiers.

1. Préparation du fichier son, enregistrement et " suppression "[3]

Précisons tout d'abord que, lors de l'enregistrement, il est impératif que l'enquêteur ait veillé à ce que chaque lecture de mot comprenne dans le même temps la lecture du nombre (adéquat !) suivie de celle du mot (adéquat lui aussi ! !) l'accompagnant. Tout manquement (erreur dans le nombre et/ou dans le mot) doit, lorsque cela est possible, conduire l'enquêteur à faire reprendre le lecteur afin de l'amener à procéder à la lecture attendue. Mais là comme ailleurs, un certain nombre d'imprévus est toujours à envisager, ces derniers n'étant d'ailleurs pas sans conséquence quant à la mise en forme du fichier son. Aussi est-il bon de rappeler ici les termes de la section 1. " Rendu des données " in *Format des rendus* (Jacques Durand, Bernard Laks, Chantal Lyche, 2002), paragraphe repris dans la révision 2004 par Julien Eychenne, Philippe Hambye et Géraldine Mallet :

> " Lors de la numérisation de la liste de mots et du texte, il est impératif de supprimer les commentaires qui précèdent et qui 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 ".

On notera que les consignes données ici concernant les listes de mots s'appliquent aussi aux listes complémentaires si les chercheurs concernés souhaitent utiliser les outils PFC.

Des informations supplémentaires ont de plus été ajoutées dans diverses versions de notre protocole, comme par exemple dans les remarques suivantes :

[1] Nous tenons à remercier Jacques Durand et Julien Eychenne pour leurs remarques et suggestions.

[2] Les consignes de rendu exposées ou rappelées dans cette notice n'ont nullement été édictées par les auteurs de cette même notice. La tâche de ces derniers n'a consisté, pour ce qui est de ces consignes, qu'à rassembler, mettre en forme et préciser des informations déjà présentes dans les conventions de PFC.

[3] Pour un rappel général des informations cf. Jacques Durand, Chantal Lyche, Bernard Laks, Mai 2002 " Protocole d'enquête ", et Jacques Durand, Bernard Laks, Chantal Lyche Mai 2002 " Format des rendus " ainsi que dans la version révisée Julien Eychenne, Philippe Hambye, Géraldine Mallet 2004.

> " … il est impératif d'épurer les fichiers de tout ce qui n'est pas liste ou texte à proprement parler (commentaires précédant la liste, remarques à la fin de la liste **ou en cours de lecture**[4], etc.)… "

Il est nécessaire d'apporter à ce propos quelques précisions. Si toute " suppression " ou " épuration " avant le début ou après la fin de la lecture même de la liste de mots reste relativement aisée, procéder à cette même opération à l'intérieur de cette " lecture " est un acte infiniment plus délicat qui nécessite d'être réalisé avec la plus extrême précaution. Aussi aura-t-on le souci d'utiliser pour cela un logiciel d'édition de fichier audio approprié[5] et de ne pratiquer ce type d'intervention que lorsque celle-ci sera à la fois nécessaire et sans risque pour le signal sonore devant être conservé. Dès lors, quand cela sera possible, tout commentaire entre deux lectures de mots pourra être supprimé. Prenons par exemple le cas d'un locuteur énonçant :

> " 1 roc euh je ne sais pas si je lis bien ce qui est écrit mais reprenez-moi si ce n'est pas ce que vous voulez 2 rat",

tout ce qui est énoncé entre les deux lectures de mots (i.e. entre " 1 roc " et " 2 rat ") pourra être enlevé, à savoir : " euh je ne sais pas si je lis bien ce qui est écrit mais reprenez-moi si ce n'est pas ce que vous voulez ".

Maintenant, il se peut encore que le locuteur insère des commentaires ou des ébauches hésitantes entre la lecture du nombre et celle du mot. L'opération de suppression est ici extrêmement sensible puisqu'elle intervient dans ce qui est censé constituer un même ensemble dont les éléments sont susceptibles d'interagir. Toutefois, dans le cas où des commentaires particulièrement abondants interféreraient entre la lecture du nombre et celle du mot qui le suit, on pourra, là encore lorsque cela ne sera pas préjudiciable à l'intégrité des données, enlever ces commentaires ou hésitations par trop marqués. Ainsi, dans l'énoncé suivant :

> " 10 euh non c'est 11 ah non je ne me suis pas trompé excusez-moi mais je crois que je vais un peu vite je vais essayer de faire attention fêtard "

on enlèvera la partie du signal (si et seulement si la configuration de ce dernier le permet) correspondant à " euh non c'est 11 ah non je ne me suis pas trompé excusez-moi mais je crois que je vais un peu vite je vais essayer de faire attention ".

Ces opérations de suppression sont particulièrement sensibles et peuvent être lourdes de conséquences quant à la fiabilité du rendu, aussi seules des personnes suffisamment averties et maîtrisant un bon logiciel d'édition de fichiers son devront procéder à de telles tâches. On évitera par ailleurs toute suppression par trop systématique et l'on n'y recourra que lorsque le " parasitage " sera réellement excédentaire, de sorte que seront laissées les hésitations simples ou légères de type " euh " ou encore les bruits de souffle, de page… Cependant, ce bruit ne devra en aucun cas être isolé dans un intervalle spécifique, et devra être "rattaché" à un intervalle contenant un mot. Par exemple, on découpera[6] :

---

[4] Le texte a été ici mis en gras par nous.

[5] Par exemple les logiciels CoolEdit (maintenant Adobe Audition), WaveLab (Steinberg), WaveStudio (Creative), … ces deux derniers étant en outre fournis avec les cartes sons Sound Blaster. Compte tenu de l'importance de ces manipulations qui exigent toute la souplesse et la fiabilité d'un logiciel spécialisé, et ce afin d'éviter toute conséquence préjudiciable, on évitera par précaution de les pratiquer à l'aide de Praat dont les fonctionnalités ne permettent pas ici d'opérer avec toute cette souplesse et fiabilité requises.

[6] Le signe % est ici employé pour représenter une frontière d'intervalle (*boundary*). Les commentaires *(bruit de page)* ne sont donnés que pour éclairer notre illustration. Ils ne doivent en aucun cas figurer dans le rendu final.

1 roc (bruit de page) % 2 rat
et non :
1 roc % (bruit de page) % 2 rat

2. Alignement sous " Praat " du signal et du texte lu[7]

Comme pour l'alignement de n'importe quel fichier de données, on procédera au moyen de la commande [TO TEXTGRID] à la création d'un fichier texte lié au fichier son ouvert ainsi que d'une tire " transcription orthographique ". Le fichier texte et la tire de transcription orthographique seront nommés conformément aux normes d'étiquetage[8].

Avant d'aborder la question même de l'alignement du signal et de sa transcription, il est au préalable nécessaire de rappeler brièvement quelques généralités quant à l'identification des fichiers et de leur(s) composante(s). Un rendu complet de la lecture de la liste de mot doit comprendre un fichier son (au format " .wav ") et un fichier TextGrid (au format " .textgrid ") ce dernier comprenant la tire de transcription orthographique de la liste de mots. On procédera à l'étiquetage de chacun des fichiers (son + TextGrid) de la même manière que celle déjà en vigueur pour les fichiers relatifs aux lectures de textes et aux conversations. La seule différence est que sera ici utilisée la lettre " **m** " pour indiquer qu'il s'agit de la liste de mots. Ainsi pour reprendre l'exemple du témoin Marie Dubois, soit " 31cmd1 "[9], le fichier son sera identifié ainsi : *31cmd1**m**w.wav*, et le fichier TextGrid de cette manière : *31cmd1**m**g.textgrid*. Enfin, la tire de transcription orthographique (où sont notés les mots précédés du chiffre : *1 roc, 2 rat...*) devra porter un label du type *31cmd1**m**_liste_mots*.

Une fois le fichier texte et la tire orthographique créés, on procédera à la segmentation de chaque ensemble nombre+mot par l'insertion d'une " marque " entre chacun de ces ensembles. Chaque intervalle délimité par ces marques pourra accueillir la transcription correspondant au signal sonore.

La transcription insérée dans chacun des intervalles correspondra de manière très limitative au seul texte lu par le locuteur, à savoir le **nombre** et le **mot** de la liste. En ce qui concerne la transcription du nombre, celui-ci ne sera suivi d'aucun point (par exemple transcrire *1 roc*, et non *1. roc*) et ce contrairement à la transcription figurant dans le texte lu par locuteur (et donné dans le protocole)[10]. En cas d'erreur de lecture (dans le nombre ou/et dans le mot) on transcrira le texte de la liste et non la réalisation effective. Par exemple, si au lieu de lire " 3 jeune ", comme il est marqué dans la liste, le locuteur réalise " 4 jeune ", on transcrira le texte devant être lu, à savoir " 3 Jeune ". De même, si le locuteur réalise " 9 nous pendrions " au lieu de " 9 nous prendrions ", on transcrira " 9 nous prendrions ". Cette transcription orthographique n'a ici aucun rôle descriptif quant à la réalisation effective par le locuteur. Sa seule fonction est de servir le repérage et la classification de ces réalisations.

---

[7] Pour une présentation de l'utilisation de " Praat " dans le cadre de PFC, cf. DELAIS – ROUSSARIE E., DURAND J., LYCHE C., MEQQORI A., TARRIER J.-M. (2002).

[8] Le lecteur pourra se référer ici à DURAND J., LAKS B., LYCHE C. (2002) ainsi que dans la révision EYCHENNE J., HAMBYE P., MALLET G. (2004).

[9] Enquête dans le département français **31** (Haute Garonne), au point d'enquête **c** (Toulouse banlieue), du témoin **md1** (Marie + Dubois + 1).

[10] Les personnes qui auraient déjà commencé ce travail de transcription en adoptant l'écriture " avec point " pourront recourir à un petit outil permettant d'extraire de leur transcription tous les " points " initiaux, outils qui sera mis très prochainement à leur disposition.

On comprendra que, dans cette optique, aucune annotation ou commentaire relatif à la performance du locuteur n'est ici à envisager (et notamment en ce qui concerne la réalisation de pause, d'hésitation, de souffle etc.).

3. Des procédures automatisées

En guise de conclusion, nous signalerons l'existence d'outils permettant d'apporter une certaine automatisation dans les tâches de segmentation et d'étiquetage des fichiers mots. En effet, souvent longues et fastidieuses, ces dernières peuvent être considérablement allégées par l'utilisation du script " PFC_Mots.praat " écrit pour le logiciel PRAAT. Ce script permet :

- de générer automatiquement un fichier TextGrid vide dont les portions correspondent aux segments inter-pauses détectés ;
- de " remplir " les portions du script précédent avec le texte correspondant au numéro et au mot lu par le locuteur.

Pour une description de cet outil ainsi que de son utilisation, le lecteur pourra se reporter à Auran & Tarrier (2004).

Références

AURAN C., TARRIER, J.-M. (2004, ce volume). " Manuel d'utilisation de l'outil d'étiquetage semi-automatique des listes de mots PFC " in Eychenne J., Mallet G. (éds.) *Bulletin PFC n°3, Du segmental au prosodique : protocoles, outils, extensions et travaux en cours,* ERSS UMR 5610 CNRS & Université de Toulouse-Le Mirail.

DELAIS – ROUSSARIE E., DURAND J., LYCHE C., MEQQORI A., TARRIER J.-M. (2002). " Transcription des données : outil et conventions " in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d'analyse,* pp 21-34. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

DURAND J., LAKS B., LYCHE C. (2002). " Format des rendus 2002 et 2003 " in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d'analyse*, pp 71-74. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

DURAND J., LYCHE C., LAKS B., (2002). " Protocole d'enquête " in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d'analyse*, pp 7-19. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

EYCHENNE J., HAMBYE P., MALLET G., (2004, ce volume). "Format des rendus" in Eychenne J., Mallet G. (éds.) *Bulletin PFC n°3, Du segmental au prosodique : protocoles, outils, extensions et travaux en cours,* ERSS UMR 5610 CNRS & Université de Toulouse-Le Mirail (à paraître).