LE FIL D'ARIANE DE LA BASE PFC

« Protocoles et méthodologie pour une base de Phonologie du Français Contemporain (PFC) »

Version 1.0 (janvier 2004)

Richard Walter

0. Introduction

Ce texte est extrait du cahier des charges de la base PFC en cours d'élaboration par Richard Walter (ingénieur d'études CNRS) et Atanas Tchobanov (ingénieur d'études Université Paris 10).

1. Site et base PFC

Janvier 2003 : le site internet PFC de présentation du projet et le travail de structuration de la base PFC sont lancés. Janvier 2004 : une autre étape est franchie avec le lancement de la convergence entre le site et la base PFC, pour qu'à l'été 2004, un seul objet concrétise le projet.

À l'heure actuelle, le site « public » (infolang.u-paris10.fr/pfc) ne délivre que les informations sur le projet. Le site « administration » (infolang.u-paris10.fr/pfc/administration) est encore en développement et permet de configurer et d'alimenter la structure de la base PFC. Ce site « administration » présente les données documentaires que nous avons pu recueillir : les équipes, les participants, la bibliographie, les enquêtes, les fiches locuteurs. Des exemples de présentation des transcriptions et des fichiers sonores sont enfin proposés.

Ce site « administration » a pour objet immédiat de :
1. Dresser un état des lieux de l'avancement du projet ;
2. Recueillir les données documentaires ;
3. Formaliser ces données par des modèles communs de description pour faciliter leurs usages finaux ;
4. Formater ces données en un format unique et transportable en différents formats ;
5. Extraire des classements et des analyses statistiques.

Ce site présentera l'avancement de ce travail, et sera remplacé à terme par la base PFC.

Cette base sera évolutive avec un corpus original et unique, homogénéisé et exploitable par différentes applications ou outils. Elle répondra à deux besoins :
1. Homogénéiser, indexer et diffuser toutes les données du projet ;
2. Exploiter ces données de multiples façons selon les besoins de l'utilisateur final.

Elle doit permettre la création de sous-corpus spécifiques pour l'application d'outils soit génériques soit spécifiques. Son système informatique doit alors favoriser l'ajout de nouveaux traitements des transcriptions et des données sonores mais doit aussi pouvoir intégrer un nouveau type d'analyse des transcriptions et/ou de nouvelles transcriptions.

Elle reposera sur quatre principes :
1. Système unique de navigation, d'interrogation et de lecture multiformat ;
2. Articulation entre la structure du corpus, les outils d'analyse et les données de ce corpus ;

3. Évolution possible du corpus et de ses outils ;
4. Implantation évolutive en et hors ligne, sous différents supports informatiques.

Elle doit répondre à quatre types d'opérations sur les données du projet :
1. Récupération & formatage ;
2. Archivage ;
3. Consultation ;
4. Exploitation.

Avec cette base, nous pourrons dépasser la consultation monolocuteur pour une exploitation à grande échelle du corpus recueilli, respectant ainsi les exigences actuelles de visibilité et de diffusion des grands corpus scientifiques.

2. Données de la base

Le projet PFC rassemble plusieurs types de données dont la base envisagée devra stabiliser les liaisons :

1. Des données hétérogènes qui composeront une bibliothèque virtuelle : introduction au projet, présentation des méthodes, protocoles (enquête, analyse, transcription), données bibliographiques, articles généraux sur le domaine, articles spécifiques sur des aspects du projet, Bulletins PFC, etc. Les données sont sous forme de fichiers avec de multiples formats (HTML, PDF, DOC, etc.).

2. Des données structurées présentant les descripteurs documentaires : équipes, participants, zones géographiques, enquêtes, locuteurs.
   Le nommage normalisé des fichiers est nécessaire mais insuffisant : il y a besoin d'une formalisation précise des descripteurs (« tel locuteur qui a tel âge et dont l'enregistrement de la conversation libre est médiocre »). Ces descripteurs donnent une représentation des données, de leur type et de leur contenu. Ils en permettent donc l'accessibilité, s'ils sont normalisés par des formulaires (phase de centralisation) et interrogeables par des requêtes (phase d'exploitation). Ils seront stockés sous forme de base de données et accessibles via une interface internet.

   3. Des données balisées : les transcriptions & les codages. Ces données sont sous forme de fichiers ASCII, avec un système normalisé de nommage, de description et de structure. Sur un même objet, elles se déclinent en trois fichiers différents, correspondant aux trois types de codages (orthographique, schwa, liaison). Ce codage est spécifique et normalisé (protocole Textgrid).

4. Des données sonores : lectures du texte et de la liste de mots, conversations libre et guidée. Par locuteur, quatre fichiers sonores, analysés et nettoyés, sont proposés. Le format d'origine de ces fichiers est WAV et le nommage des fichiers est standardisé à l'identique du 'nommage' des transcriptions.
   Les fichiers sonores seront archivés au format WAV original mais accessibles en ligne en MP3. Une exploitation fine des enregistrements sonores sera toujours possible avec le format WAV initial.

3. Système de la base

Le corpus de la base PFC est composé de données avec des relations structurées et basées sur un protocole commun (système classique de base de données relationnelles). Sur ce corpus, s'appliquent deux systèmes :

1. Système de requêtes pour les données documentaires :

- Requêtes simples sur les données textuelles (« Lire les articles ou les textes de présentation pour telle enquête, tel participant ou telle problématique »).
- Requêtes complexes multicritères :
    - Définir une catégorie de locuteurs et un besoin spécifique (« Écouter la lecture du texte ou de tel mot par des locuteurs de plus de 60 ans ou de Toulouse ») ;
    - Écouter les fichiers sonores ou les parties de fichier sonore ;
    - Lire et analyser en parallèle les différentes transcriptions.

Un lien automatique des réponses vers les transcriptions et les fichiers sonores est rendu possible grâce à la normalisation du nommage des fichiers et à la formalisation des descripteurs.

2. Système de fichiers pour la création de sous-corpus :

Un besoin d'analyse phonétique doit provoquer la création d'un sous-corpus sonore et textuel :

- Définir un phénomène phonétique (« Écouter toutes les prononciations du mot pâte », « Écouter tous les Schwas dans tel contexte ») ;
- Analyse des fichiers de transcriptions par un outil de type moteur de recherche et classement des réponses dans un ordre déterminé ;
- Extraction des données sonores correspondantes par un outil spécifique ;
- Regrouper les transcriptions et leurs données documentaires ;
- Constitution du corpus spécifique ;
- Lire, écouter et analyser en parallèle ce corpus.

La lecture et l'analyse des résultats se feront sous trois formats (données documentaires, transcriptions, extraits sonores) mais par une interface unique consultable par navigateur internet.

Il va de soi que les deux systèmes peuvent se combiner pour répondre à des besoins comme celui-ci : « Écouter telle liaison par des locuteurs de plus de 60 ans ou de Toulouse »

4. Outils de la base

La base PFC reposera sur un système de requêtes : recherche, mise à jour (ajout, modification, suppression), extraction de données et de corpus.

D'autres outils font la richesse du projet PFC :

1. Des outils de codage : outils utilisés en amont pour fabriquer les fichiers de transcription et de codage ;
2. Des outils d'exploitation : outils utilisés en aval pour analyser des sous-corpus de la base, créés pour des besoins spécifiques.

Les premiers sont ou seront diffusés par le site PFC, suivant des modalités de diffusion à définir ; les seconds devront pouvoir s'appliquer sur des données textuelles HTML ou ASCII et/ou sonores WAV ou MP3.

5. Base PFC en ligne

La base PFC sera accessible via un navigateur internet et diffusera les types d'information suivants :

- Présentation du projet et informations générales (déjà disponibles sur le site PFC) ;
- Pages locales des équipes ;
- Production et références du projet ;
- Exemples significatifs du corpus et de son exploitation, en accès public ;

- Outils de codage et d'exploitation développés pour des besoins spécifiques du projet ;
- Données recueillies lors des enquêtes et transcrites numériquement.

Une politique de droits d'accès et d'utilisation gérera la consultation de cette base avec un module de « gestion de sessions » par mot de passe, permettant, pour l'utilisateur, de garder la trace de ses consultations et de ses interrogations.

La base PFC sera stockée à différents endroits :
1. Un serveur à l'université Paris 10 ;
2. Un certain nombre de sites miroirs propriétés des équipes du projet. Ces sites respecteront un protocole de mise à jour via le serveur de l'université Paris 10 et auront des droits d'accès spécifiques.

Elle sera mise à jour par deux procédures :
1. Des procédures de référencement des données via des formulaires permettront aux participants du projet de centraliser et de tenir à jour leurs données ;
2. Des procédures de mise à jour permettront de synchroniser les différentes localisations de la base sur les sites miroirs.

6. Référencement de la base

Cette base sera accessible aux participants du projet pour qu'ils puissent introduire leurs données, les modifier et les exploiter. Des formulaires de saisie (ajout, correction, suppression) seront opérationnels mais accessibles par un mot de passe. Les formulaires ont fait l'objet d'un consensus pour les critères de réponses en quantité et en formulation. Un temps de saisie sera donc nécessaire pour chaque participant du projet, tout comme de revenir sur les données de terrain et de les synthétiser.

Ces formulaires proposeront une saisie souple des données documentaires. Celles-ci sont stockées dans une base de données et consultation avec le même format unique (HTML pour les données documentaires, HTML et ASCII [Textgrid] pour les transcriptions). Pour les consulter ou les stocker pour un usage propre, il suffira alors pour l'utilisateur de sauvegarder sur son ordinateur la page affichée par le navigateur utilisé.

Les formulaires sont accessibles de partout et à n'importe quel moment. Les participants au projet peuvent donc à tout moment saisir ou modifier leurs propres données.

7. Accessibilité de la base

Afin de préserver la confidentialité des travaux en cours et de respecter les us et coutumes en matière de diffusion de corpus électronique, il n'y aura jamais d'accès à l'intégralité des transcriptions et des données sonores, hormis pour la direction du projet. La granularité des rôles dans le projet PFC déterminera les différents accès possibles aux données de la base :

1. Niveau « consultation » qui limite la consultation à l'intégralité du sous-corpus introduit dans la base par l'utilisateur et, suite à une requête, à des contextes limités pour le reste du corpus. Ce niveau concernera le « participant » au projet.
2. Niveau « enquête » qui limite la consultation à l'ensemble du corpus d'une enquête et à des contextes limités pour les autres enquêtes. Ce niveau concernera les responsables d'enquête ou d'équipe du projet.
3. Niveau « développement informatique » qui autorise la création de sous-corpus à partir du corpus entier de la base. Ce niveau concernera les correspondants locaux des équipes développant des outils PFC.

4. Niveau « administration » qui ouvre l'ensemble du corpus et des fonctionnalités. Ce niveau sera réservé à la direction du projet et aux développeurs de la base.

Chaque responsable d'enquête, d'équipe ou du projet pourra déléguer des droits de consultation, fixes ou temporaires.

8. Calendrier de la synergie

Courant février 2004, les formulaires de référencement seront disponibles par mot de passe à l'ensemble des participants. Dès l'annonce de cette disponibilité, chacun devra demander par courrier électronique un mot de passe qui donnera des « droits d'écriture » spécifiques. Ces droits permettront de mettre à jour les données propres de chaque participant et de chaque équipe. De par la demande d'un mot de passe, ceux-ci prennent la responsabilité de l'actualisation et de la pertinence des informations les concernant. Par ailleurs, plusieurs personnes pourront avoir les « droits d'écriture » pour la même équipe ou la même enquête.

Fin mars 2004, une première interface graphique d'interrogation de la base sera proposée avec les modules de traitements statistiques de toutes les données. Après validation de cette interface, les participants du projet pourront interroger, suivant leur droit d'accès, l'ensemble des données présentes dans la base.

La structure finale de la base sera présentée en juillet 2004. Elle devra continuer à être mise à jour par les nouvelles enquêtes terminées et se prêtera aux exploitations prévues pour des usages spécifiques. Une nouvelle version du site PFC sera alors proposée, mise « au goût du jour » des usages de l'internet et avec une interface graphique pour l'instant inexistante.

Cette montée en puissance de la base PFC montre qu'il faut aller encore plus vers la convergence et la stabilisation des données et des outils. L'objectif est toujours de créer un corpus réellement utilisable et visible auprès de la communauté scientifique concernée.