Présentation et avancée du projet Traitement de corpus oraux en français. Description et comparaison de productions langagières (interactions entre adulte, entre adulte et enfant)

Emmanuelle Canut

Présentation et avancée du projet « Traitement de corpus oraux en français. Description et comparaison de productions langagières (interactions entre adulte, entre adulte et enfant) »

Emmanuelle Canut, ATILF-Nancy Université
Emmanuelle.Canut@univ-nancy2.fr
www.emmanuelle.canut.neuf.fr

Objectifs généraux du projet : un grand corpus pour quoi faire en recherche ?

Le projet cherche à rassembler de nombreux corpus (plusieurs millions de mots) pour diverses raisons :

- archiver les données dans un format moderne en vue d'une meilleure conservation et d'un gain de place (archivage sous forme numérique et non plus sur des cassettes), et en vue d'une meilleure « visualisation » des corpus (pouvoir lire la transcription et écouter le son en même temps, ce qu'on appelle l'alignement texte-son) ;
- archiver les données pour aller plus vite et plus loin dans les analyses selon un effet accumulatif de la recherche (enrichir les analyses existantes en les confrontant à d'autres corpus existants sans avoir à recommencer toute la procédure de recueil et de transcription des données) ;
- travailler sur un nombre de données statistiquement pertinent ;
- mettre à disposition (dans une certaine mesure) les corpus en vue d'un partage des données dans la communauté scientifique. Outre la base de données anglo-saxonne CHILDES, qui comporte assez peu de corpus en français, il n'existe pas de « plateforme » en France qui soit accessible aux chercheurs.

Mais plus précisément, archiver de nombreux corpus permettrait de mettre au jour des processus généraux d'apprentissage grâce à :

- la description linguistique (aspects lexicaux, syntaxiques et pragmatiques) et la comparaison de nombreuses pratiques langagières dans des situations variées d'interaction entre un adulte et un enfant ;
- la comparaison (du point de vue de leurs caractéristiques linguistiques) entre les verbalisations d'adultes dans des situations de communication ordinaire avec d'autres adultes (récits, conversation, explication, etc.) et les verbalisations des adultes adressées à de jeunes enfants (moins de 7 ans) dans des situations d'apprentissage ou de vie quotidienne.

Mais comment faire ?

Avant de pouvoir réaliser ces objectifs de recherche, il y a toute une partie « technique » à mettre en place :

- des outils (machine avec grande capacité de stockage, matériel pour numériser les cassettes, logiciels divers) ;
- des moyens « humains » (pour numériser, aligner, transcrire les données) ;

- des collaborations diverses (en particulier avec des informaticiens, des chercheurs et ingénieurs spécialistes en Traitement Automatique du Langage) pour évaluer les besoins et avancer dans la réalisation technique du projet.

Mais il y aussi une nécessaire réflexion en amont : quelles modalités de transcription faut-il adopter ? Comment archiver les données ? Seront-elles en accès libre (travail autour des aspects éthiques, juridiques…) ? Faut-il adopter une normalisation internationale ? Etc.

Carte d'identité du projet

1- Liste des participants de l'ATILF (Nancy Université)

- *Emmanuelle Canut*, responsable du projet : enseignante-chercheure, Sciences du langage. Associée à l'équipe CRALOE-CALIPSO de l'Université Paris 3.
- *Jeanne-Marie Debaisieux*, responsable du projet : enseignante-chercheure, Sciences du langage. Associée à l'équipe DELIC de l'Université de Provence.
- *Evelyne Jacquey*, responsable du projet : chercheure CNRS (axe Lexique). Membre du projet interne DIXEM.
- *Virginie André* : enseignante-chercheure, Sciences du langage ;
- *Isabelle Clément et Christiane Jadelot*, ITA (Ingénieurs, Techniciens et personnels Administratifs du CNRS) – Equipe Bases Textuelles et TAL.

D'autres personnes du laboratoire ont également collaboré sous des formes diverses (numérisations de données, stages linguistique et informatique, transcriptions) : *Françoise Weiss* et *Josette Frecher* (ITA), *Magali Husianycia* (doctorante SDL), *Suzanne Dombret, Jérome Marchiset* et *Youma Sow* (master SDL).

2- Les collaborations

- Pour la constitution et l'archivage de corpus oraux : Martine Vertalier et Luiggi Sansonnetti : Equipes CRALOE – CALIPSO (EA 1324) et SYLED (EA2290), Université Paris 3 (archivage de corpus en acquisition du langage).
- Pour l'exploitation informatique de corpus oraux : Jean Véronis et Christophe Benzitoun : DELIC, Université Aix-Marseille (utilisation des logiciels *Transcriber* et *Contextes*).
- Pour l'analyse informatique des données : les membres du projet DIXEM, ATILF (constitution et/ou capitalisation d'outils d'enrichissements de corpus et de leurs interrogations).
- Pour la normalisation des corpus en vue d'un archivage « publique » (selon la norme internationale TEI) : Mathieu Quignard : LORIA (Equipe « TALARIS ») et Bertrand Gaiffe : ATILF (Equipe « LEXIQUE » et « Ressources et normalisation »)

3- les soutiens au projet

Le projet a démarré en septembre 2005 et a bénéficié d'un soutien (technique et financier) du laboratoire ATILF depuis cette date. Au cours de l'année 2005-2006, sur le plan logistique, l'ATILF a mis à disposition :

- une salle de travail avec une machine ayant une grande capacité de stockage et un double disque, un écran 19 pouces et l'appareillage nécessaire pour réaliser des numérisations (casques audio, magnétophone) ;
- un compte spécial dans le domaine ATILF permettant d'assurer des sauvegardes régulières des données numérisées et transcrites.

La direction du laboratoire a également accepté que des ITA consacrent 150h à l'alignement texte-son des données (le projet avait été soutenu initialement à hauteur de 50 h par ITA).

L'ATILF a également financé des vacations pour la numérisation : 3000€ en 2005, 2006 et 2007 (correspondant à 540h de numérisation) et la rémunération de 3 stages de 3 mois chacun (2700 €).

En 2007, le projet bénéficie en supplément d'une autre source de financement (via le Contrat Plan Etat Région) qui permet notamment de réaliser d'autres transcriptions et d'embaucher un ingénieur pour travailler sur la normalisation des données.

Avancées de la réflexion

Au cours de l'année universitaire 2005-2006, parallèlement au travail « technique » (logistique), une réflexion a été amorcée sur la constitution, l'informatisation et la diffusion des données. Cette réflexion a porté sur :

- le choix des conventions de transcription, parmi celles existantes : il fallait tenir compte de la « lisibilité » des corpus par la communauté scientifique (conventions identiques pour tous) et en même temps de la spécificité des données et des analyses envisagées (choix d'une transcription orthographique assez « nue », signes particuliers liés aux spécificités du langage enfantin, etc.). Mais il fallait aussi tenir compte du fait que des transcriptions allaient être désormais réalisées sous le logiciel *Transcriber* (avec un alignement texte-son) ce qui impliquait également l'adoption d'autres normes[1] ;
- les critères à retenir pour constituer un archivage des données en fonction des caractéristiques des corpus (réalisation de fiches de suivi et de fiches descriptives pour chaque corpus selon une codification unique[2]) ;
- les modifications à apporter aux anciens modèles d'autorisation pour l'enregistrement et la diffusion des données, en se référant au « Guide des bonnes pratiques » (DGLFLF version 2005) et aux conseils des personnes compétentes dans l'ATILF pour la mise en forme « juridique »[3].

Au cours de l'année universitaire 2006-2007, en collaboration avec des informaticiens (Matthieu Quignard et Bertrand Gaiffe), une autre phase de la réflexion a été entamée, liée davantage à la normalisation des données (selon une norme internationale : la TEI[4]) pour rendre l'archivage des données visible à l'extérieur :

- comment passer de transcriptions faites avec des logiciels d'alignement texte-son (comme *Transcriber* et *Praat*) vers un format TEI et inversement ?
- comment passer d'un format TEI vers un format plus « agréable » à lire (comme un *pdf*) ?

Ce travail axé sur l'informatisation des données a été mené conjointement à une autre réflexion plus « linguistique » :

- Quels sont les outils (logiciels) en accès libre qui nous permettent de faire des analyses quantitatives de données orales ?
- Peut-on typologiser pour faire un archivage des données ? Quels critères ?[5]

[1] Les conventions dorénavant adoptées pour faire des transcriptions sur un traitement de texte (type *Word*) ou plus spécifiquement pour faire un alignement texte-son avec le logiciel *Transcriber* sont disponibles et envoyées à toute personne qui le souhaite. Contactez : Emmanuelle.Canut@univ-nancy2.fr.

[2] On trouvera en annexe les fiches de suivi du projet. Pour un exemple de fiche descriptive, voir ci-après l'article de Magali Husianycia.

[3] Les nouveaux modèles d'autorisation sont envoyés à toute personne qui le souhaite. Contactez : Emmanuelle.Canut@univ-nancy2.fr.

[4] Une explication de la TEI est donnée ci-après par Jérome Marchiset.

[5] Ces deux questions sont abordées ci-après par Magali Husianycia.

Les différentes phases du projet

La première année (2005-2006), nous avons pu réaliser le travail suivant :

- sauvegarde sous forme numérique d'environ 80 h d'enregistrements sur cassettes audio ;
- harmonisation des conventions de transcription (basées en partie sur celles du DELIC en raison de leur adéquation avec l'utilisation du logiciel d'alignement texte-son *Transcriber* et du logiciel de concordances *Contextes*) ;
- formation à la transcription et au logiciel *Transcriber* de quatre personnes ITA du laboratoire. Depuis janvier 2006, une centaine de corpus ont été transcrits ;
- répertoire de l'ensemble des corpus qui constitueront la base de données du projet sous forme de fiche de suivi.
- stage informatique (3 mois) ciblé sur l'archivage des données (assuré par Jérome Marchiset). L'objectif était de gérer de manière contrôlée le stockage et la maintenance des données d'ors et déjà numérisées et transcrites et les mettre en forme dans un format pivot (étude des conventions TEI) afin d'assurer l'éventuelle mise à disposition des données auprès de la communauté scientifique dans un format partagé.

La deuxième année (2006-2007), nous avons pu réaliser le travail suivant :

- sauvegarde sous forme numérique d'environ 250h d'enregistrements sur cassettes audio ;
- stage linguistique (assuré par Magali Husianycia). L'objectif était de : 1) répertorier des outils à disposition sur internet permettant de faire des requêtes ou des relevés quantitatifs, de les tester sur des corpus oraux numérisés et transcrits, d'indiquer ce que ces outils permettent ou non de faire, les qualités et les défauts de chacun en fonction de la perspective d'analyse proposée dans le projet ; 2) de vérifier la validité des fiches descriptives, notamment pour la répartition des corpus en genre de discours, sous-genre, type de parole, type d'interaction…

Actuellement, le travail suit son cours :

- Une réflexion plus poussée sur la TEI pour les corpus oraux dans le cadre du CPER (Contrat Plan Etat Région) et sur les données à anonymiser (lesquelles ? Comment ?).
- Un deuxième stage linguistique, correspondant au mémoire de Master 2 de Suzanne Dombret, qui a pour objectif :
    - le repérage de l'élément « relatif qui » dans différents corpus oraux numérisés de français parlé et le classement de cette construction selon des caractéristiques linguistiques (syntaxique et sémantique) à définir (réalisation d'une « typologie ») ;
    - le repérage de ce même élément dans des corpus transversaux et longitudinaux d'interaction adulte-enfant avec établissement d'un classement utilisant les mêmes critères que pour les productions de l'adulte et tenant compte de leur ordre d'apparition dans le temps.
- La poursuite des transcriptions des données (alignement texte-son) par trois personnes ITA du laboratoire, ainsi que via de nouvelles vacations et des étudiants de licence 3 en Sciences du langage.
- La préparation de deux autres sujets de mémoire de master 2 Sciences du langage sous la direction de Emmanuelle Canut.

ANNEXE

1- FICHE DE SUIVI DU TRAITEMENT DES DONNES NUMERISEES (Mai 2007)
Corpus d'interactions individuelles adulte-enfant
Responsable : Emmanuelle Canut

Corpus numérisés	**Nombre**	**Durée totale**	**Âges des enfants**	**Transcriptions disponibles**	**Alignement texte-son**	**Anonymisation**	**Disponibilité**
Uniques	141	2159mn12s = 35h59mn12s	2;3 à 5;10 (+ 1 enfant de 1;10 et 1 de 6;5)	Transcriptions inexistantes (33), partielles non vérifiées en version Word (94) ou papier (7), ou transcriptions non vérifiées intégrales (7)	Néant	Néant	Accès restreint (104) ; Accès libre (37)
Longitudinaux	63 corpus longitudinaux représentant au total 918 corpus	8387mn16s = 139h47mn16s	2;10 à 7;6 (+ 1 enfant à partir de 1;11 et 1 enfant au-delà de 9 ans)	Transcriptions inexistantes (5), partielles ou intégrales non vérifiées en version Word ou papier (758), ou transcriptions en cours de vérification (158)	En cours pour 16 corpus longitudinaux (= 158 corpus)	Néant	Tous en accès restreint sauf un en accès privé
TOTAL	**1059**	**175h46mn31s = 1 M 400 000 mots**					

2- FICHE DE SUIVI DU TRAITEMENT DES DONNES NUMERISEES (Mai 2007)
Corpus de français parlé
Responsables : Virginie André et Jeanne-Marie Debaisieux

Corpus numérisés	**Nombre**	**Durée totale**	**Transcriptions disponibles**	**Alignement texte-son**	**Anonymisation**	**Disponibilité**
- *Parole publique/privée* - *Relation pro/perso* - *Entretien/ conversation/ interview*	Environ 250 corpus (dont 16 réunions en milieu pro)	Environ 156h (dont 32h réunions pro) = 1 M 250 000 mots	Transcriptions inexistantes ou partielles non vérifiées en version Word ou transcriptions non vérifiées intégrales	En cours	Néant (seulement sur version Word)	Accès restreint (corpus perso et réunions pro) ou accès libre (corpus étudiants si anonymisation)