Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement des Corpus Oraux en Français)

Emmanuelle Canut, Virginie André, Bertrand Gaiffe

Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français)

1. Origine et objectifs du projet : quelles données pour quelles recherches ?

Le projet TCOF (Traitement de Corpus Oraux en Français) que nous présentons ici a pour ambition de collecter un nombre conséquent de données orales, des dialogues pour la plupart, en vue d'une diffusion internationale libre et gratuite. Ce projet, qui s'inscrit dans une réflexion actuelle sur la collecte et le traitement des données en français parlé, est rattaché à l'Unité Mixte de Recherche 7118 CNRS et Nancy Université : ATILF (Analyse et Traitement Informatique de la Langue Française). Il comporte actuellement 21 heures de corpus alignés texte-son (version Transcriber et TEI) qui peuvent être téléchargeables gratuitement : http://www.cnrtl.fr/corpus/tcof/

1.1. Contexte scientifique de la recherche sur les données orales

Dans le contexte de la recherche française, les données recueillies par les chercheurs en sciences humaines ne sont pas toujours librement et totalement à la disposition de l'ensemble de la communauté scientifique. Il existe ainsi, en Sciences du langage, une multitude de données orales et écrites sur le français mais leur accès est très inégalement assuré :

- En ce qui concerne l'écrit, des bases de données textuelles existent et sont consultables. Par exemple, *Frantext* est un corpus à dominante littéraire constitué de textes français qui s'échelonnent du XVIe au XXIe siècle. Le CNRTL (Centre National de Ressources Textuelles et Lexicales) regroupe des corpus informatisés comme les deux années des éditions intégrales du quotidien régional *l'Est républicain*. Internet constitue également en soi une base de données.
- La situation est plus problématique en ce qui concerne les données orales. Ces dernières sont éparpillées et rarement consultables par des personnes extérieures à la recherche locale. « On peut poser qu'il y a sans doute entre quatre ou cinq millions de mots effectivement disponibles mais l'absence de coordination rend l'exploitation de l'ensemble impossible » (Debaisieux, 2005). Les données sont, en outre, fortement hétérogènes dans la mesure où le recueil et l'analyse dépendent de l'objectif d'étude, de l'orientation épistémologique et de la connaissance des outils à disposition de chaque chercheur (Habert, 2000). En conséquence, la mutualisation des connaissances issues de chacune des exploitations pour les corpus oraux reste relativement exceptionnelle, ce qui rend difficile le cumul des analyses (pour une tentative de synthèse voir la revue *Lidil*, 31, 2005 : « Corpus oraux et diversité des approches », pour un inventaire, nécessairement incomplet, des corpus oraux en France voir Cappeau et Seijido, 2005). Par exemple, au sein du laboratoire ATILF, si nous faisons la somme des données orales en possession de chacun des chercheurs, nous obtenons un total proche de 2 millions de mots.

Certains pays européens (Allemagne, Angleterre, Espagne, Portugal, Italie) ont dépassé ces difficultés et ont pu constituer un corpus de référence contenant aussi bien de l'écrit que de l'oral :

- le British National Corpus (BNC) : cette banque de données compte 100 millions de mots enrichis par des annotations morphosyntaxiques.
- le Corpus de Référence de l'Espagnol Actuel (CREA) : cette banque de données compte actuellement 100 millions de mots et devrait encore s'en enrichir de vingt-cinq millions. Elle présente une grande variété d'extraits écrits et oraux, produits dans tous les pays hispanophones depuis 1975.
- le corpus de référence allemand COSMAS II (Corpus Search, Management and Analysis system : Institut *Für Deutsche Sprache* à Mannheim).
- le Corpus de Référence du Portugais Contemporain (CRPC) : cette banque de données orales et écrites compte actuellement 80 millions de mots.
- Le corpus du néerlandais (Corpus Gesproken Nederlands) et le projet associé de corpus de l'écrit (Dutch Language Corpus Initiative) : environ 9 millions de mots pour le corpus oral.

Par ailleurs, concernant plus spécifiquement le langage des enfants, le CHILDES (*Child Language Data Exchange System*) est une des rares bases de données existantes d'interactions verbales spontanées entre les jeunes enfants et leurs parents, leurs enseignants ou des camarades[1]. Les différents programmes informatiques associés à la base de données (pour une analyse automatique dans le format CHAT[2]) et l'ensemble des corpus (audios et/ou vidéos) sont disponibles gratuitement et téléchargeables dans leur intégralité. Il existe actuellement plus de 200 corpus transcrits se rapportant à une trentaine de langues différentes. Il peut s'agir de corpus longitudinaux (un même enfant ou de très petits groupes d'enfants enregistrés sur plusieurs années) ou de corpus transversaux (des groupes d'enfants enregistrés à quelques périodes). Pour le français, on répertorie une quinzaine de corpus longitudinaux (enregistrements dans le cadre familial d'enfants d'âges compris entre 1 an et 11 ans, sur des périodes allant de quelques mois à plusieurs années).

Prenant conscience du retard de la France dans le développement des ressources et la constitution d'une banque de données textuelles informatisée, notamment pour la langue parlée, la communauté universitaire a entamé depuis 2000 une réflexion de fond, en particulier pour ce qui concerne la constitution et l'hébergement de corpus :

- Création d'un « Guide des bonnes pratiques » qui fait le point sur les aspects déontologiques, juridiques et techniques du recueil et de l'analyse de données orales (*Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*, Délégation Générale à la Langue Française et aux Langues de France, 2006).
- Regroupement sur une même base internet des enregistrements concernant, entre autres, le français et ses variétés et les langues de France. Ce projet, sous l'égide de la DGLFLF (Délégation Générale à la Langue Française et aux Langues de France, via le conseil scientifique de l'Observatoire des pratiques linguistiques), s'inscrit dans le cadre d'un accord entre le CNRS et le Ministère de la culture pour prolonger et rendre cohérent sur le long terme le programme *Corpus de la parole* qui donne la priorité aux ressources orales.

[1] Base de données fondée par B. MacWhinney et C. Snow en 1984. Le projet implique des milliers de chercheurs à travers le monde, et des milliers d'articles ont été publiés à partir de données de la base : http://childes.psy.cmu.edu/.

[2] Instructions et normes destinées à assurer une standardisation des procédures de transcription et de codage phonologique, morphologique et en actes de parole.

- Création par le CNRS d'un Centre de Ressources pour la Description de l'Oral (CRDO) pour la conservation et la diffusion de corpus oraux.
- Création de bases de données informatisées :
  - ⇨ ESLO (« Enquête sociolinguistique à Orléans », laboratoire CORAL, Université d'Orléans) ;
  - ⇨ PFC (« Phonologie du Français Contemporain ») ;
  - ⇨ CFPP2000 (« Corpus de Français Parlé Parisien des années 2000 », Université Paris 3-Sorbonne Nouvelle)
  - ⇨ CLAPI (« Corpus de LAngue Parlée en Interaction », Unité Mixte de Recherche CNRS : ICAR, Université Lyon 2)[3].
  - ⇨ LEONARD (« Acquisition du langage et grammaticalisation », Projet ANR rattaché à la base CLAPI)[4].

1.2. A l'origine du projet, des ambitions, des besoins et des objectifs…

Le projet TCOF a démarré en septembre 2005 et a depuis toujours bénéficié d'un soutien financier et logistique du laboratoire CNRS ATILF[5]. A l'origine du projet se trouvent des linguistes[6] qui ont accumulé depuis les années 1990 un grand nombre de données orales (des enregistrements audio) dans deux domaines de recherche distincts : l'analyse syntaxique des productions orales d'adultes en français parlé et l'analyse des interactions entre adultes et enfants âgés de 2 à 7 ans en linguistique de l'acquisition. Ces chercheures ont souhaité mettre en commun l'ensemble de ces données pour des exploitations à la fois individuelles et collectives. La convergence entre les deux orientations s'est faite autour d'un projet portant sur la description et la comparaison des productions langagières, notamment du point de vue de l'usage des formes linguistiques chez les locuteurs adultes (s'adressant à un autre adulte ou à un enfant) et chez les locuteurs enfants, en lien avec les interactions verbales.

Une équipe s'est peu à peu constituée, élargissant notamment le domaine des recherches à l'analyse des interactions verbales entre adultes en situation de travail[7], intégrant un autre chercheur en analyse descriptive du français[8] et faisant appel à d'autres compétences, en particulier celles de spécialistes en linguistique informatique[9].

---

[3] Voir la présentation de CLAPI dans ce même numéro.

[4] Il s'agit d'un des rares projets de « plateforme » en France d'enregistrements d'enfants monolingues entre 12 mois et 3 ans, filmés dans leur famille. Généralement, les chercheurs en acquisition disposent d'une base de données personnelle. Par exemple, D. Bassano et son équipe ont constitué une base de données de « corpus français de productions langagières précoces » comportant, entre autres, deux corpus longitudinaux et un corpus transversal codé selon un système approprié aux objectifs de leur recherche, mais l'accès n'est pas public (voir Bassano, 2005, p. 66-67 pour une description).

[5] L'ATILF assure la logistique du projet (mise à disposition d'un ordinateur de grande capacité de stockage, sauvegarde des données, appareils et logiciels nécessaires à la numérisation) et a financé des vacations et des stages pour la numérisation, la transcription et l'anonymisation des données, à hauteur de 3000€ minimum par an.

[6] Il s'agit de Jeanne-Marie Debaisieux, d'Emmanuelle Canut et Martine Vertalier (Nancy Université et Université Paris 3).

[7] Il s'agit des données et de la recherche de Virginie André (Nancy Université).

[8] Christophe Benzitoun (Nancy Université), qui travaille sur les liens entre lexique et grammaire en mettant en évidence la proportion des phénomènes de figement (par exemple : contre/par contre – cause /à cause de) ou de contrainte (par exemple : causer un préjudice, un accident…tous termes à visée négative) massivement représentés à l'oral non planifié.

[9] Bertrand Gaiffe, Etienne Petitjean et Evelyne Jacquey (CNRS ATILF), en collaboration avec d'autres chercheurs du laboratoire CNRS LORIA (équipe TALARIS avec Mathieu Quignard dans un premier temps puis équipe PAROLE avec Christophe Cerisara, Dominique Fohr et Odile Mella, voir 3.3. dans cet article).

Le projet est lié à une volonté de partage des ressources et à des besoins de recherche, en particulier celui d'affiner la description de la langue orale, celui de mettre au jour les processus interactionnels qui régissent les pratiques langagières, ou encore les processus interactionnels d'apprentissage au cours de la période d'acquisition du langage.

De ce fait, l'objectif est double :

- Pouvoir rassembler de nombreuses initiatives actuellement éparpillées, ce qui signifie :
    1. archiver les données dans un format unifié et explicite en vue d'une conservation pérenne, d'une projection plus simple vers les autres formats de données et de développer des outils permettant de visualiser l'ensemble des informations disponibles (notamment pour combiner simultanément l'accès à la transcription et à l'écoute du son) ;
    2. 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) ;
    3. travailler sur un nombre de données statistiquement pertinent.

- Faire une étude syntaxique et/ou interactionnelle des productions orales à partir d'un grand nombre de corpus, en particulier :
    1. la description linguistique (aspects lexicaux, syntaxiques) et pragmatique des pratiques langagières ;
    2. la comparaison (du point de vue de leurs caractéristiques linguistiques) entre des verbalisations d'adultes dans des situations de communication ordinaire avec d'autres adultes (récits, conversation, explication, etc.) et des verbalisations d'adultes s'adressant à de jeunes enfants (moins de 7 ans) ;
    3. la comparaison entre les productions linguistiques des adultes et celles des jeunes enfants (mise en lien entre diversité du répertoire linguistique proposé et développement langagier) ;
    4. la possibilité de faire des recherches transversales sur la langue, s'appuyant à la fois sur des corpus écrits et sur des corpus oraux.

Nous avons dès le départ envisagé la diffusion des corpus avec une mise à disposition libre et gratuite en vue d'un partage au sein de la communauté scientifique. Dès lors, avant même d'avoir envisagé une analyse des données, se sont posées des questions liées à la conservation, à l'archivage et à la visibilité informatique des données, aussi bien sur le plan technique que juridique. Cependant, la question de l'analyse est restée forcément présente, puisqu'il a fallu réfléchir à la mise en commun de données assez différentes (dialogues entre adultes et interaction adulte-enfant).

2. Etapes de construction du projet

Avant d'aboutir à une version finalisée des corpus, la constitution d'un pôle « technique » a été nécessaire :

- 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 a aussi en amont une réflexion sur la constitution, l'informatisation et la diffusion des données : quelles modalités de transcription faut-il adopter ? Comment archiver les données ? Faut-il anonymiser le son et la transcription et, si oui, comment (en lien avec les aspects éthiques et juridiques pour un accès libre) ? Existe-t-il des normes internationales et quelle est celle qui convient le mieux aux buts que nous poursuivons ?

A l'heure actuelle, la plupart des transcriptions d'oral sont dans le format de l'outil utilisé pour effectuer la transcription (Praat, Transcriber, Elan ou Clan dans la majorité des cas) ou, lorsque le fichier sonore n'a pas été conservé, dans le format utilisé par un logiciel de traitement de texte. A titre d'exemple, le CRDO autorise des dépôts de corpus dont la transcription est au format texte utf8, pdf, rtf ou praat.

Parmi les choix de format indépendant d'un outil donné et correctement décrits, voire normalisé, la TEI (Text Encoding Initiative) bien qu'envisagée dans certains projets (par exemple, le Dutch Spoken Corpus qui a finalement définit un format spécifique) n'arrive pas véritablement à s'imposer.

De même, en ce qui concerne les métadonnées, au-delà du format Dublin Core, les projets adoptent les formats les plus divers : de OLAC (une forme de Dublin Core qualifié dédiée aux données linguistiques) à des formats très riches, propres à un projet (cf. Childes) ou normalisés tels qu'IMDI (Isle Metadata Initiative).

Nos choix actuels et les perspectives d'évolution que nous envisageons font l'objet des paragraphes suivants.

2.1. Recueil des données

Les données primaires sont constituées d'enregistrements audio recueillis par des chercheurs de Nancy[10], des étudiants en Licence et Master Sciences du langage à l'Université Nancy 2 et à l'université Paris 3. Il s'agit de données orales recueillies dans des contextes aussi naturels que possibles :

- Des interactions entre adultes comportant :
  1. des données sollicitées
     - ⇨ des entretiens dans lesquels au moins deux locuteurs sont engagés dans des récits de vie, d'évènements ou d'expériences, ou dans des explications sur un savoir faire professionnel ou technique ;
     - ⇨ des conversations à bâtons rompus ou portant sur des thématiques spécifiques.
  2. des données non sollicitées
     - ⇨ des situations publiques ou professionnelles : réunions publiques, activités professionnelles diverses.

[10] V. André, E. Canut et J.M. Debaisieux. Voir aussi la « Présentation » de *Verbum*, (Debaisieux, Bertin et Husianycia, à paraître).

Cet échantillonnage est susceptible d'être affiné en fonction des données recueillies notamment par les étudiants dans les années à venir.

- Des dialogues entre un adulte et un enfant de moins de sept ans comportant des données longitudinales (de quelques mois à plusieurs années) et des enregistrements uniques :
  - ⇨ des conversations libres (thématiques variées) ;
  - ⇨ des narrations à partir de livres illustrés[11].

Une des premières phases du projet a consisté en une importante logistique de numérisation des données (des enregistrements sur cassettes se sont poursuivis jusqu'en 2007-2008)[12] et l'archivage sur un même espace dédié (avec constitution d'un répertoire de suivi).

Projet TCOF
Corpus adulte-enfant : 176 h numérisées
Corpus adulte-adulte : 250 h numérisées

2.2. Transcription des données

Au fil de la réflexion, nous avons élaboré plusieurs versions des conventions de transcriptions et profondément modifié le format de nos données. Sans vouloir entrer précisément ici dans les enjeux de la transcription (Ochs 1979, Blanche-Benveniste 1997, Mondada 2000, Debaisieux 2005), les choix opérés par les acteurs du projet tiennent compte de la diversité de leurs approches et des exploitations envisagées (notamment syntaxique et interactionnelle). Ainsi, nous sommes passés de conventions de transcriptions divergentes (utilisées sur traitement de texte) à des conventions communes. Ce travail d'harmonisation, amorcé dès la première année, devait tenir compte de la « lisibilité » du format utilisé pour la communauté scientifique et en même temps de la spécificité des données et des analyses envisagées. Nous avons notamment fait le choix d'une transcription orthographique standard assez « nue » avec peu de symboles afin d'éviter de surcharger le texte[13]. Seuls quelques signes particuliers, liés aux spécificités du langage enfantin, ont été ajoutés dans les corpus adulte-enfant.

Le choix de réaliser un alignement texte-son avec le logiciel *Transcriber* impliquait également d'adopter d'autres normes liées à l'utilisation des balises d'annotations internes au logiciel[14]. Le tableau, ci-dessous, présente les principaux symboles[15] introduits dans la

[11] Les situations de co-narration avec le support de livres illustrés peuvent être considérées comme semi-expérimentales dans la mesure où le chercheur peut être impliqué dans le dialogue avec une visée particulière d'apprentissage.
[12] Utilisation du logiciel *Wavelab* pour numérisation dans le format *.wav*. Ce format a été choisi parce qu'il est le seul compatible avec *Transcriber*.
[13] L'Université d'été « Transcription de langue parlée. Aspects théoriques, pratiques et technologiques » de Perpignan en juin 2005 avait déjà conclu en préconisant des transcriptions nues afin de faciliter le partage et la mutualisation des données (Bilger, 2008).
[14] *Transcriber* étant un logiciel libre, il ne faisait pas obstacle à la diffusion de nos corpus. Néanmoins, nous avons complété les fichiers générés par *Transcriber* à l'aide de traitements spécifiques, en fonction de nos nécessités d'analyse. Nous avons ainsi ajouté des modules supplémentaires comme la numérotation des tours de parole et les chevauchements de plus de deux locuteurs (voir *infra* 3.3.).
[15] L'ensemble des conventions de transcription est disponible sur le site : http://www.cnrtl.fr/corpus/tcof/TCOFConventions.pdf

transcription, soit comme caractère, soit comme balise *Transcriber*. Ces conventions sont exemplifiées dans l'interface reproduite ci-dessous.

RECAPITULATIF DES SYMBOLES DE TRANSCRIPTION	
{…}	Commentaires (balise *Transcriber*)
[…]	Prononciations particulières notées avec l'alphabet phonétique SAMPA (balise *Transcriber*)
(…)	Variantes graphiques indécidables
+	Pauses
///	Pauses très longues
=	Liaison non standard remarquable
/…, …/	Hésitations entre transcription, multi-transcription
…-	Amorces
*	Syllabe incompréhensible
***	Suite de syllabes incompréhensibles
###	Passage enregistré non transcrit
$$$	Coupure de l'enregistrement

Figure 1 : Interface de Transcriber

La facilité d'utilisation de ce logiciel a également influencé notre choix. La transcription est réalisée par les chercheurs (notamment lorsqu'il s'agit de leurs propres données), des étudiants de Sciences du langage dans le cadre de la constitution de dossiers[16], des vacataires

16 Les étudiants de Licence 2ème année ont un cours intitulé « Constitution de corpus », assuré par V. André et J.-M. Debaisieux, dans lequel ils découvrent la linguistique de corpus et sont amenés à recueillir leur propre corpus (reversé au projet TCOF avec leur accord). Les étudiants de Licence 3ème année transcrivent et vérifient des corpus dans un cours de linguistique de l'acquisition (E. Canut)

et certains membres du laboratoire[17]. Les données transcrites sont systématiquement vérifiées par les responsables du projet avant leur diffusion.

2.3. Les aspects juridiques

Une partie du travail préparatoire a porté sur les aspects juridiques de la mise à disposition de données orales et nous a amené à modifier les anciens modèles d'autorisation concernant l'enregistrement, le traitement, l'exploitation et la diffusion des données. Nous appuyant sur les recommandations publiées dans le *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux* et sur les conseils des personnes du CNRS compétentes en la matière[18], nous avons proposé de nouvelles autorisations (incluant une cession de droits d'auteur pour les dépositaires de données) et nous sommes mis d'accord sur le principe d'une anonymisation des données graphiques et sonores interdisant l'identification des locuteurs. Suite à ces réflexions, nous avons instauré :

- un accès restreint (aux seuls chercheurs de l'équipe) aux données susceptibles de renseigner l'identité des personnes comme le nom de famille ou la date de naissance ;
- un accès libre sur le site, avec une anonymisation de la transcription et du signal sonore des noms de villes ou de sociétés, des noms de famille… renseignant sur l'identité des individus et sur les organisations. Par exemple, dans la transcription, nous avons systématiquement remplacé les noms de familles par ¤N¤ et nous avons remplacé ces noms de famille par un « bip »[19] dans le fichier son. Les symboles représentant la version anonyme des noms de familles, toponymes, sociétés ou associations sont numérotés selon leur ordre d'apparition dans l'enregistrement.

Ainsi, les corpus déposés sur la plateforme qui sont téléchargeables sont uniquement des versions anonymisées[20].

2.4. Constitution des métadonnées

Les corpus sont accompagnés d'une fiche descriptive (aujourd'hui en cours d'informatisation) comportant les métadonnées. Plusieurs versions de fiches signalétiques ont été nécessaires avant de se fixer sur une codification homogène et une présentation cohérente. La question du contenu et de la pertinence de certains champs décrivant le corpus (genre de discours/typologie textuelle/canal de communication) a notamment été très discutée. Bien entendu, au-delà de nos besoins propres, nous avons enrichi notre réflexion à l'aide des propositions d'IMDI, mais aussi des métadonnées utilisées dans les autres projets (BNC[21], PFC, CRFP, C-ORAL-ROM, etc. Les informations contenues dans cette fiche sont regroupées dans différentes catégories :

[17] La direction du laboratoire ayant acceptée que du personnel ITA (Ingénieurs, Techniciens et Administratifs) consacrent 150h par an à la transcription, cinq personnes ont été formées à la transcription et à l'utilisation du logiciel *Transcriber* (Gisèle Cagne, Isabelle Clément, Josette Frecher, Christiane Jadelot et Françoise Weiss). Il en fût de même pour des étudiantes vacataires (Stéphanie Houin, Youma Sow et Cécile Desse).

[18] Il s'agit notamment des compétences liées à la déclaration des données à la CNIL (Commission Nationale de l'Informatique et des Libertés).

[19] Ces manipulations se font avec le logiciel libre *Audacity*.

[20] Les besoins d'anonymisation des corpus adulte-enfant sont nettement moins importants que dans les interactions entre adultes.

[21] Voir à ce sujet les articles de Crowdy (1993) et Lee (2001).

- des informations générales (titre du corpus, type d'accès, logiciel d'alignement, type d'anonymisation, nombre de locuteurs, relations entre les participants, modalités de recueils des données, cadre situationnel, canal de communication, genre de discours, type d'association entre les corpus, documents annexes/artefacts, résumé…)
- des informations concernant l'enregistrement (notamment : nom de la personne qui a enregistré, type d'autorisation, validité juridique de l'autorisation, support original, type d'enregistreur numérique, qualité du son, format et taille du fichier son, durée, lieu de l'enregistrement, conditions d'enregistrement…)
- des informations sur la transcription (nom des transcripteurs, nom des réviseurs, format, conventions, nombre de mots…)
- des informations sur les locuteurs (étiquette dans la transcription, nombre de tours de parole, âge, sexe, niveau d'étude, profession, rôle dans l'interaction, degré de planification du discours, statut du français, autre(s) langue(s) d'usage, lieu de naissance, lieu de résidence, appartenance régionale dominante…)
- des informations sur les travaux et les publications réalisés à partir du corpus
- le mode de référencement du corpus (à citer par l'utilisateur : nom, nom du responsable, titre du projet, laboratoire, lien vers la description de TCOF)

Sur la plateforme, les corpus peuvent être interrogés à partir des métadonnées suivantes : nombre de locuteurs ; type de corpus ; cadre situationnel ; genre de discours ; âge du locuteur ; sexe du locuteur ; type d'association entre les corpus ; canal de communication ; statut du français ; autre(s) langue(s) d'usage ; degré de planification ; appartenance régionale dominante ; résumé.

La réflexion autour du contenu des métadonnées a également amené à se poser la question de l'échantillonnage des données. L'objectif serait à la fois de proposer une typologie basée sur des critères nettement identifiables et d'identifier les secteurs devant faire l'objet d'enregistrements et de transcriptions en priorité. Cette étape semble indispensable à terme pour que nos descriptions puissent prendre en compte la dimension contrastive, préalable indispensable à toute entreprise dont la finalité est de mettre en évidence le caractère polymorphe du langage.

CARTE D'IDENTITE PROJET TCOF

http://www.cnrtl.fr/corpus/tcof/

Le projet « Traitement de Corpus Oraux en Français » (TCOF) est né au départ de la volonté de conserver des corpus oraux constitués dans les années 1980-90 à des fins de recherches personnelles. L'équipe constituée au sein du laboratoire ATILF (UMR CNRS 7118) a élaboré l'architecture d'une base de données de corpus alignés texte/son qu'elle a progressivement enrichie. L'équipe met à disposition de la communauté scientifique ces ressources, au fur et à mesure du traitement des données.

Données du corpus TCOF

Le corpus comporte ainsi des enregistrements en situations réelles de deux catégories :
- *des interactions entre adultes* : conversations, entretiens, réunions publiques/professionnelles, débats, plaidoyer, consultations.

Le projet met l'accent sur la variation des situations d'énonciation et sur la diversité discursive.
- *des interactions entre adulte et enfant* : conversations, narrations avec support du livre illustré.
Les corpus sont uniques ou longitudinaux (de quelques mois à plusieurs années).
Les enregistrements sont de tailles diverses : de 5' à 45' ou plus.

Les corpus disponibles sont transcrits, anonymisés et ont été vérifiés à plusieurs reprises. La disponibilité des corpus est fonction de critères juridiques : seuls sont disponibles les corpus pour lesquels les autorisations de diffusion sont valides et pour lesquels l'anonymisation a été réalisée.

CORPUS ADULTES
Ressources potentielles
250 h d'enregistrements numérisés dont :
- 180 h d'enregistrements soit non transcrits, soit transcrits, non anonymisés
- 70 h de corpus non diffusables (problèmes juridiques)

Ressources actuellement téléchargeables
- 15 h d'enregistrements représentant 72 transcriptions

CORPUS ADULTE-ENFANT
Ressources potentielles
176 h d'enregistrements numérisés dont :
- 36 h d'enregistrements (141 corpus uniques) transcrits, non vérifiés, non anonymisés.
  => Enfants âgés de 2 ans 3 mois à 5 ans 10 mois
- 140 h d'enregistrements (63 corpus longitudinaux, représentant 918 corpus) transcrits, non vérifiés, non anonymisés
  => Enfants âgés de 2 ans 10 mois à 7 ans 6 mois

Ressources actuellement téléchargeables
- 6 h d'enregistrement (5 corpus longitudinaux, représentant au total 62 corpus)
  => Enfants âgés de 3 ans 10 mois à 5 ans 10 mois

3. Traitement informatique des corpus et projets associés

L'ultime étape de création d'une plateforme (hébergée par l'INIST - INstitut de l'Information Scientifique et Technique - CNRS) pour la diffusion des données et des métadonnées a impliqué (et implique encore) de répondre à plusieurs problématiques informatiques[22]. Même si le traitement informatique des corpus a nécessité une réflexion approfondie depuis le début du projet (André, Benzitoun, Canut, *et al.*, à paraître), nous présenterons ici uniquement le choix du codage en TEI, l'interrogation automatique des données et les projets associés à TCOF, menés en partenariat avec des chercheurs en informatique.

3.1. Le codage en TEI

[22] Ce travail fait notamment appel aux compétences de Bertrand Gaiffe, Etienne Petitjean, Kamel Nebhi et Cyril Pestel (ATILF).

Plusieurs éléments nous ont poussés à choisir le format de codage TEI (*Text Encoding Inititative*) pour l'ensemble de nos corpus. Ainsi, pour que les données soient dans un format partageable et si possible pérenne, et pour que les corpus puissent être enrichis *a posteriori*, il a été fait le choix de les normaliser selon la norme internationale TEI. De plus, cela nous permet :

- un même codage des fiches signalétiques (TEI Header) et des transcriptions (le passage du format *Trs* de *Transcriber* vers TEI a été automatisé ainsi que le passage de TEI vers PDF afin de garantir aux chercheurs une bonne lisibilité des données et un format correspondant à leurs besoins) ;
- une intégration des méta-données et des données dans un même fichier, avec une entête inamovible. Cette intégration nous garantit la localisation de l'ensemble des éléments : fichier son, transcription, métadonnées (ces trois éléments étaient auparavant distincts et simplement liés par le même nom de fichier) ;
- un travail beaucoup plus fin sur l'annotation ;
- une interrogation plus facile des corpus à partir des métadonnées.

Il s'agit cependant d'un travail de longue haleine et ce format ne sera appliqué aux corpus que progressivement. Enfin, le format TEI étant largement documenté, l'échange éventuel de corpus sera facilité.

3.2. L'interrogation des données

Une exploitation des données par recherche de concordances est possible sur le site de diffusion grâce à l'intégration d'un concordancier, réalisé spécialement pour le projet. De nombreux concordanciers sont actuellement librement disponibles, certains disposant d'une expressivité en termes de requêtes tout à fait satisfaisant (le logiciel *AntConc*, par exemple). Pourtant, force est de constater qu'il n'existe pas à notre connaissance de logiciels libres permettant d'effectuer des concordances sur des corpus alignés texte/son. Actuellement, les membres du projet TCOF utilisent le logiciel *Contextes* développé à l'Université de Provence par Jean Véronis et mis à la disposition de l'équipe depuis 2007 mais il ne s'agit pas d'un logiciel libre et les licences qui sont fournies avec le logiciel sont monopostes. Le concordancier développé dans le cadre du projet TCOF, intitulé *JConc*, a pour objectif de tirer pleinement partie des corpus transcrits dans le cadre du projet, et au-delà, de rendre un grand service à la communauté des chercheurs travaillant sur l'oral[23].

Le concordancier que nous avons développé est un logiciel qui effectue des recherches de chaînes de caractères et/ou de mots et les retourne accompagnés de leur contexte. Ce logiciel est plus particulièrement orienté vers le travail sur des transcriptions d'oral, même s'il fonctionne également sur des corpus écrits. Par exemple, il met également en relation le texte et le son, ce qui permet l'écoute de l'extrait sonore correspondant à un passage textuel. Ce logiciel est proche de *Contextes* et écrit en Java.

Pour résumer, *JConc* possède notamment les fonctionnalités suivantes[24] :

- recherche d'un mot (ou d'une chaîne de caractères)
- utilisation d'expressions régulières
- ouvrir en entrée un fichier Transcriber ou un texte brut en ISO-8859-1

[23] Nicolas Guth et Vincent Poutissou ont développé cet outil d'abord dans le cadre de leur stage de Licence Informatique à l'Université Nancy 1 puis dans le cadre de vacations engagées par le laboratoire ATILF.

[24] Des nombreuses fonctionnalités sont ajoutées au fur et à mesure des tests que nous réalisons de ce logiciel.

- contraindre les contextes droit et gauche
- affichage :
    - du numéro de l'occurrence
    - du nom du fichier dans lequel se trouve l'occurrence
    - des contextes
    - de l'identifiant du locuteur
    - du début et de la fin du segment en secondes ainsi qu'en caractères
    - du texte complet d'où est extraite l'occurrence et mise en valeur dans le texte
- écoute du passage audio correspondant à :
    - la lecture du segment dans lequel se trouve l'occurrence
    - la lecture en continu à partir du segment
    - la lecture complète du fichier son
- recherche par locuteur
- exportation de la recherche dans un fichier lisible par un tableur (fichier tabulaire)
- calcul de statistiques basiques (du type fréquence de l'occurrence selon le locuteur par rapport à l'ensemble du corpus…)

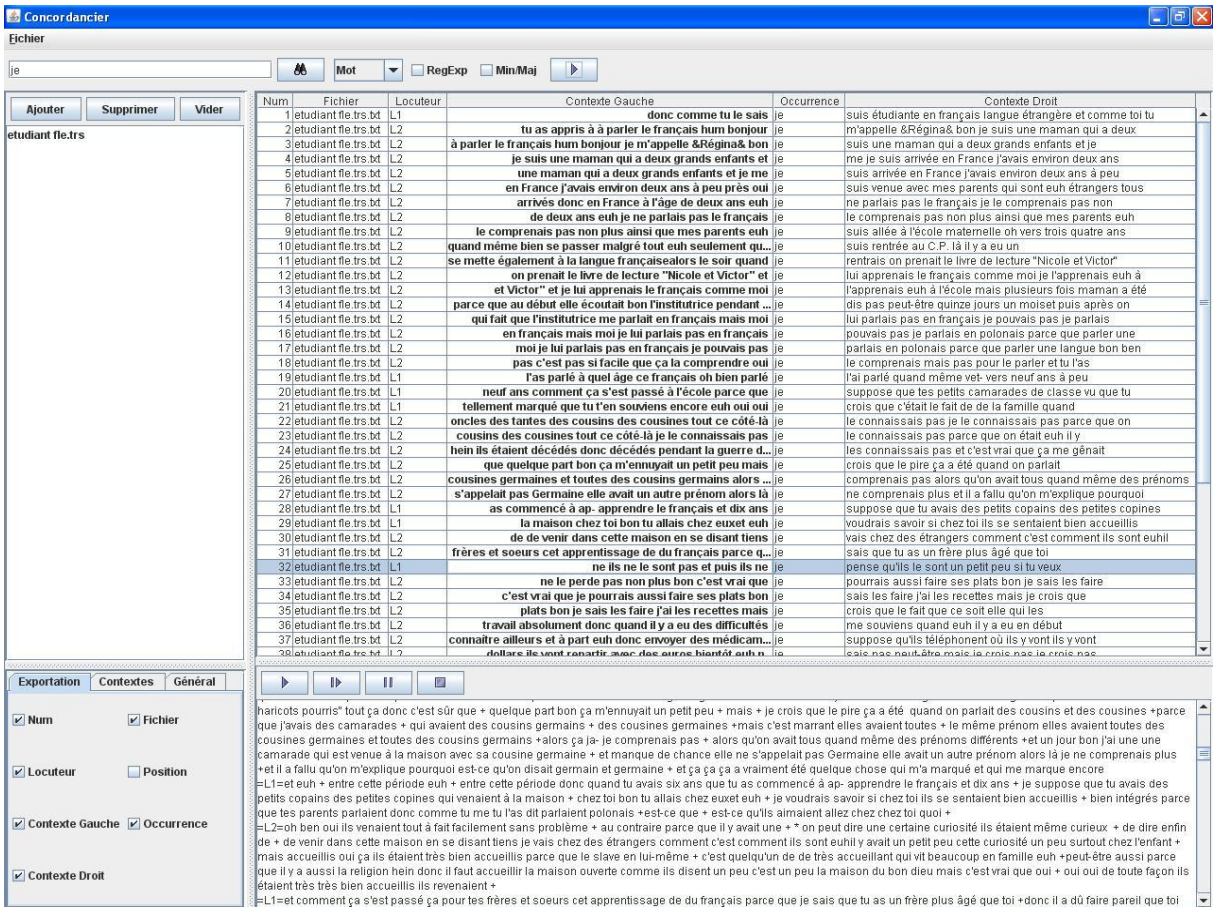

Figure 2 : Interface de JConc

Lorsque la conception de ce logiciel sera entièrement terminée[25], il sera disponible sur le CNRTL, libre et téléchargeable gratuitement.

[25] Au moment de la rédaction de cet article, les principales fonctionnalités de *JConc* seront disponibles dans les semaines à venir.

3.2. Projets informatiques associés

Pour mener à bien ce projet TCOF, et afin d'initier et de mettre en place un réseau de coopérations entre les communautés scientifiques s'intéressant respectivement aux ressources linguistiques, au traitement automatique de la parole, et au dialogue, des collaborations se sont créées avec le LORIA (laboratoire LOrrain de Recherche en Informatique et ses Applications, UMR 7503). Ces collaborations sont nées d'intérêts communs concernant la parole et des différents besoins de traitement informatique des données de TCOF. Nous avons collaboré avec deux équipes du LORIA. Tout d'abord, nous nous sommes associés avec l'équipe TALARIS (Traitement Automatique des Langues : Représentations, Inférences et Sémantique). Cette équipe s'intéresse notamment à la linguistique computationnelle et principalement aux objets de recherche qui mobilisent la sémantique et l'inférence en développant des modèles d'analyse du discours. Ensuite, nous nous sommes tournés vers l'équipe PAROLE, spécialisée dans l'analyse, la perception et la reconnaissance automatique de la parole. Leurs travaux en traitement automatique du signal cherchent à extraire le sens, à analyser et à renforcer la structure acoustique de la parole.

Avec l'équipe TALARIS[26], nous avons poursuivi la réflexion concernant les formats de codage des corpus oraux. Dans ce cadre, afin de garantir aux utilisateurs une souplesse dans le choix de leurs formats de codage, des outils de conversion ont été créés afin de passer du format Transcriber au format TEI (et inversement). Pour améliorer la lisibilité des transcriptions, notamment lors de l'impression papier de ces dernières, nous avons également élaboré un outil permettant de convertir les formats Transcriber et TEI en fichier PDF. Une exportation en SMIL (Synchronized Multimedia Integration Language) a aussi été réalisée afin de visualiser d'une autre façon l'enregistrement et sa transcription. En outre, une partie du CPER MISN TALC (Contrat Plan Etat Région / Modélisation, Informations et Systèmes Numériques / Traitement Automatique des Langues et des Connaissances) a été consacrée à l'élaboration d'un module supplémentaire à intégrer dans le logiciel *Transcriber*. *Transcriber* permet de formaliser et de baliser les chevauchements de parole mais seulement lorsque deux locuteurs sont impliqués. Le module réalisé dans le cadre cette collaboration permet de gérer des chevauchements lorsque plus de deux locuteurs parlent en même temps.

Avec l'équipe PAROLE (dans le cadre de l'opération intitulée ALIGNE du CPER MISN TALC), nous avons cherché à résoudre de façon semi-automatique l'alignement texte/son des données transcrites dans TCOF. Un des objectifs est d'étudier la possibilité d'utiliser les outils stochastiques de traitement automatique de la parole afin de faciliter et de réduire le coût humain du travail d'annotation, de création de corpus et de ressources linguistiques. Cette équipe maîtrise la technologie permettant d'aligner du texte et du son de façon automatique voire de transcrire de façon quasi-automatique. Cela n'est cependant réalisable que dans des conditions très contraintes : le signal doit être pré-segmenté en fichiers courts (quelques minutes maximum), tous les mots de vocabulaire doivent être connus par le logiciel, le signal sonore doit être de très bonne qualité, sans aucun bruit, et la parole enregistrée ne doit pas être une parole spontanée, c'est-à-dire contenir des traces de la construction du discours (telles que des amorces de mot, des hésitations, des répétitions ou encore des chevauchements). Les outils standards sont donc inutilisables pour des données telles que celles du projet TCOF.

[26] Représentée par Matthieu Quignard.

C'est pourquoi, l'équipe PAROLE a créé le logiciel libre *JTrans*[27] (Cerisara, Mella, Fohr, 2009), outil d'alignement texte-parole permettant de faciliter et de réduire le temps d'alignement. A partir d'un fichier audio contenant de la parole enregistrée (discours, interview, …) et d'un fichier texte contenant une transcription textuelle de ce qui est dit dans le fichier audio, *JTrans* produit un alignement entre le texte et la parole. Ce logiciel rencontre certes des difficultés pour réaliser l'alignement des corpus qui s'éloignent des conditions idéales, c'est-à-dire les corpus avec plus de deux locuteurs, avec des locuteurs qui ont des accents régionaux prononcés, avec du lexique spécifique ou encore avec les corpus d'enfants (voix faibles, murmures impossibles à aligner). Néanmoins, même si des retouches manuelles sont encore nécessaires, le logiciel *JTrans*, améliore le temps d'alignement et rend le travail moins fastidieux.

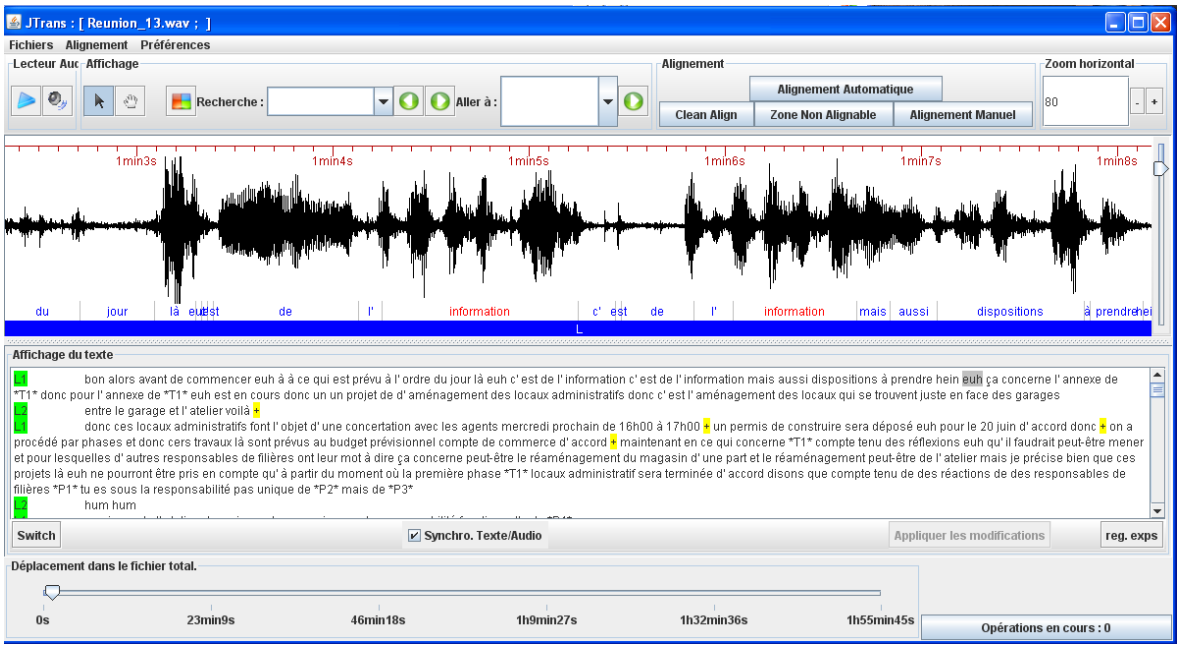

Figure 3 : Interface de JTrans

A plus ou moins long terme, le travail de l'équipe PAROLE concernant l'alignement au phonème près (réalisé par *JTrans*) dans des conditions réelles de bruitage et avec de multiples locuteurs aura une incidence sur l'anonymisation puisque cela permettra de repérer précisément le segment à anonymiser. A partir de là, nous pourrons remplacer ce segment par une courbe mélodique identique mais non identifiable.

Ces collaborations se poursuivent à l'heure actuelle…

Références bibliographiques

ANDRÉ, V., BENZITOUN, C., CANUT, E. (*et al.*), à paraître, « Traitement informatique de données orales : quels outils pour quelles analyses ? », Actes du colloque

[27] Téléchargeable sur : http://www.loria.fr/~cerisara/jtrans/index.html

AnaLogiQual 2008 Logiciels pour l'analyse qualitative : innovations techniques et sociales, Luxembourg, 21-22 octobre 2008.
BASSANO, D., 2005, « Production naturelle précoce et acquisition du langage. L'exemple du développement des noms », Lidil, 31, p. 61-84.
BAUDE, O. (Ed.), 2006, *Corpus oraux, guide des bonnes pratiques*, Paris : Editions du CNRS.
BILGER, M. (Ed.), 2008, *Données orales, les enjeux de la transcription*, Cahiers de l'Université de Perpignan, n°37.
BLANCHE-BENVENISTE C., 1997, *Approches de la langue parlée en français*. Paris : Editions Ophrys.
CAPPEAU, P. et SEIJIDO M., 2005, *Inventaire des corpus oraux en langue française*, document téléchargeable à l'adresse : www.dglflf.culture.gouv.fr.
CERISARA, C., MELLA, O., FOHR, D., 2009, « JTrans: an open-source software for semi-automatic text-to-speech alignment », In *INTERSPEECH-2009*, 1823-1826.
CROWDY, S., 1993, Spoken Corpus Design, In *Literary and Linguistic Computing*, 8-4, p. 259-266.
DEBAISIEUX, J.-M., 2005, « Les corpus oraux : Situation, exploitation linguistique, bilan et perspectives », *De la linguistique de corpus à la relation « partie/tout » : varia*, *Scolia*, 19, p. 9-40.
DEBAISIEUX, J.-M., BERTIN, T., HUSIANYCIA M. (Eds), à paraître, *Verbum*, 2008/n°3.
Equipe DELIC, 2004, « Autour du Corpus de référence du français parlé ». *Recherches sur le français parlé*, 18, p. 11-42.
HABERT, B., 2000, « Des corpus représentatifs : de quoi, pour quoi, comment ? ». In HABERT B, *Instruments et ressources électroniques pour le français*, Paris, Gap, OPHRYS.
LEE, D., 2001, « Genres, registers, text types and styles: clarifying the concepts and navigating a path through the BNC Jungle », In *Language Learning and Technology*, 5-3, available online at http://llt.msu.edu/vol5num3/lee/default.html
LIDIL, 2005, n°31, « Corpus oraux et diversité des approches ».
MONDADA L., 2000, « Les effets théoriques des pratiques de transcription ». *LINX*, 42, p.131-149.
OCHS, E., 1979, « Transcription as theory ». *In* OCHS E., SCHIEFFELIN B. (Eds), *Developmental Pragmatics*. New York : Academy Press, p.43-72.