Traitement informatique de données orales : quels outils pour quelles analyses ?

Virginie André, Maître de conférences

Université de Nancy 2

Christophe Benzitoun, Maître de conférences

Université de Nancy 2

Emmanuelle Canut, Maître de conférences

Université de Nancy 2

Jeanne-Marie Debaisieux, Maître de conférences

Université de Nancy 2

Bertrand Gaiffe, Chargé de recherche

Université de Nancy 2

Evelyne Jacquey, Chargée de recherche

Université de Nancy 2

Résumé

Dans cet article, nous présentons le projet TCOF (Traitement de corpus oraux en français) dont l'objectif est la pérennisation, la collecte et le traitement des données de français parlé. Nous mettons plus particulièrement l'accent sur la façon dont les outils informatiques, utilisés de manière conventionnelle ou détournés de leur usage habituel (logiciel d'alignement texte-son, concordanciers, etc.), font avancer le travail d'archivage, de codage et d'analyse qualitative de nos données en vue d'une diffusion internationale libre et gratuite.

RECHERCHES QUALITATIVES – Hors Série – numéro 9 – pp. 131-150.
LOGICIELS POUR L'ANALYSE QUALITATIVE: INNOVATIONS TECHNIQUES ET SOCIALES
ISSN 1715-8702 - http://www.recherche-qualitative.qc.ca/Revue.html

Mots clés
CORPUS ORAUX, TRANSCRIPTION ASSISTÉE PAR ORDINATEUR, MÉTADONNÉES ET CODAGE, ANALYSE LINGUISTIQUE QUALITATIVE, MUTUALISATION ET CUMUL DES CONNAISSANCES.

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 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, tel que Frantext, corpus à dominante littéraire constitué de textes français qui s'échelonnent du XVI^{e} au XXI^{e} siècle. Le Centre national de ressources textuelles et lexicales (CNRTL) regroupe également des corpus informatisés comme les deux années des éditions intégrales du quotidien régional *l'Est républicain*. Internet constitue 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. 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 Savelli (2005), pour un inventaire des corpus oraux en France voir Cappeau & Seijido (2005)).

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'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.

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 (Bilger, 2000; Habert, Nazarenko & Salem, 1997), 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 (Baude, 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 Délégation générale à la langue française et aux langues de France (DGLFLF), 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;
- 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 base de données informatisées : projet ESLO (Enquête sociolinguistique à Orléans, laboratoire CORAL, Université d'Orléans), projet international PFC (Phonologie du français contemporain), projet CLAPI (Corpus de langue parlée en interaction », Unité mixte de recherche CNRS : ICAR, Université Lyon 2).

Nous présenterons ici le projet TCOF (Traitement de corpus oraux en français), rattaché à l'Unité mixte de recherche CNRS - Nancy Université : Analyse et traitement informatique de la langue française (ATILF), qui est développé dans le contexte nancéien et qui illustre cette avancée de la réflexion sur la collecte et le traitement des données de français parlé (Canut, 2008). Nous mettrons l'accent sur la façon dont les outils informatiques, utilisés de manière conventionnelle ou détournés de leur usage habituel (logiciel d'alignement texte-son, concordanciers, etc.), font avancer le travail

d'archivage, de codage et d'analyse qualitative de nos données en vue d'une diffusion internationale libre et gratuite.

Origine et objectifs du projet Traitement de corpus oraux (TCOF)

À l'origine du projet se trouvent des linguistes qui ont accumulé depuis les années 1990 un grand nombre de données orales (des enregistrements audio) dans des domaines de recherche distincts : syntaxe, sociolinguistique, didactique et linguistique de l'acquisition. La convergence entre les différentes 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 et enfants, en lien avec les interactions verbales. Le projet a démarré en septembre 2005 et a depuis bénéficié d'un soutien financier et logistique du laboratoire ATILF. Il a un double objectif de recherche :

- la description linguistique, aspects lexicaux, syntaxiques, pragmatiques et interactionnels, de pratiques langagières dans des situations d'interaction entre adultes et d'interaction entre adultes et enfants au cours de l'apprentissage du langage;
- la comparaison, du point de vue de leurs caractéristiques linguistiques, entre les verbalisations entre 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.

L'étude doit permettre de mener des analyses descriptives sur un grand nombre de corpus de français parlé pour affiner la description syntaxique de la langue orale (Benzitoun, Campione, Deulofeu, Henry, Teston, Valli & Véronis, à paraître; Véronis, 2000). Elle doit permettre de rendre compte de phénomènes d'acquisition comme l'évolution du langage de l'enfant (augmentation et diversité du répertoire linguistique) et les caractéristiques linguistiques du langage de l'adulte adressé à l'enfant.

Nous avons dès le départ envisagé la diffusion des données sans restriction (mise à disposition libre et gratuite pour la communauté scientifique). Ainsi, avant même de pouvoir envisager une analyse des données, la question de la conservation des données, de leur archivage et de leur pérennité s'est posée, d'autant que leur visibilité pouvait être accrue grâce au traitement informatique. À la fin de la chaîne, et c'est la principale réflexion que nous menons actuellement, se pose le problème du croisement possible des analyses (sur le plan méthodologique en particulier mais aussi au niveau de l'analyse linguistique des données) et donc de leur mise en commun.

Constitution des données primaires

Il s'agit de données orales authentiques (enregistrées dans des situations réelles) : conversations entre adultes (récits divers, réunions de travail en entreprise, entretiens…), entre adulte et enfant de moins de 7 ans (conversations libres, narrations à partir de livres illustrés).

Le matériel qui a été utilisé pour procéder aux enregistrements est classique : magnétophone à cassettes avant l'ère du numérique, enregistreurs numériques depuis deux ou trois ans. Il a donc fallu, dans un premier temps, procéder à l'harmonisation du son en numérisant toutes les cassettes (son Wav, avec le logiciel Wavelab) et, dans un deuxième temps, archiver le tout sur un même espace (machine dédiée).

La réflexion a porté sur les aspects juridiques liés au recueil des données – notamment en ce qui concerne les autorisations – et a pris en compte les positions publiées dans le *Guide des bonnes pratiques* (Baude, 2006). Sur ce point, même si nous étions en possession d'autorisations écrites des locuteurs, nous nous sommes mis d'accord sur le principe d'une anonymisation des données graphiques et sonores interdisant l'identification des locuteurs. Pour ce faire, nous avons instauré :

- un accès restreint (aux seuls chercheurs de l'équipe) des données susceptibles de renseigner l'identité des personnes comme le nom de famille ou la date de naissance;
- une anonymisation de la transcription et du son (masquage par du bruit) des noms de villes ou de sociétés, des noms de famille… renseignant sur l'identité des personnes. Par exemple, dans la transcription, nous avons systématiquement remplacé les noms de familles par ¤P¤ et dans le fichier son nous avons remplacé ces noms de famille par un « bip ».

Constitution des données secondaires

La transcription

Au fil de la réflexion et de l'usage généralisé des technologies informatiques, nous avons élaboré plusieurs versions des conventions de transcriptions et profondément modifié le format de nos données. Ainsi, nous sommes passés de conventions de transcriptions divergentes (adulte-adulte, adulte-enfant, réunions de travail) à des conventions communes, moyennant quelques informations supplémentaires spécifiques. Nous reproduisons ici deux exemples de corpus transcrits ainsi que les symboles utilisés dans les transcriptions. Les transcriptions sont faites en orthographe standard, sans trucage orthographique (voir Tableau 1).

Nous sommes également passés de transcriptions sous format Microsoft Word ou papier vers des transcriptions alignées texte-son (points de synchronisation entre la transcription et le fichier son) avec le logiciel Transcriber (voir Figure 1).

Le passage à la transcription assistée par ordinateur a eu un impact important sur la fiabilité et le temps de transcription, notamment grâce à l'intégration dans une interface simple et conviviale de tous les modules utiles. Transcriber étant un logiciel libre, il ne fait pas obstacle à la diffusion de nos corpus (Tableau 2).

Nous avons par ailleurs 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 les chevauchements de plus de deux locuteurs et la numérotation des tours de parole.

Nous avons également, grâce à une collaboration avec le Laboratoire Lorrain de recherche en informatique et ses applications (LORIA) (UMR CNRS, Nancy), développé un logiciel en Java comparable à Transcriber (logiciel JTRANS). Il permet un transfert semi-automatique de transcriptions sous format Microsoft Word vers des transcriptions alignées texte-son. Ce logiciel, en cours de test, permettra un alignement rapide, précis et fiable des transcriptions existantes et permettra de traiter de façon rapide les différents états des transcriptions rassemblées depuis le début du projet (voir Figures 2 et 3).

Le format de codage

Une fiche documentaire accompagne chaque fichier son et sa transcription. Sa mise au point a nécessité une phase de standardisation et d'uniformisation assez longue. La fiche décrit ce qu'on appelle classiquement les « méta-données » relatives au corpus considéré (participants à la conversation, titre du corpus, responsable(s) du projet, nom du transcripteur, matériel d'enregistrement utilisé, date d'enregistrement, pour n'en citer que quelques-unes) (voir Tableau 3).

D'un point de vue pratique, la séparation entre méta-données d'un coté et fichier transcrit de l'autre, peut poser problème :

- d'une part, pour chaque corpus, il faut s'assurer de garder le lien entre sa fiche documentaire, son fichier son et sa transcription (pour l'instant ce lien repose sur une convention de nommage des fichiers et leur présence dans un même répertoire);

Tableau 1
Exemples d'application des conventions de transcription

Corpus adulte-enfant	**Corpus adulte**
VERONIQUE • là je vois la poupée avec un nounours là	L1 • je sais pas elle a dit qu'on trouverait un moyen
ADULTE • tu crois que c'est un nounours ça ? • c'est bien plus gros qu'un nounours	L2 • ah bon sûrement parce que moi je peux pas prendre de voiture
VERONIQUE • non c'est un Babar *[pron=baba]*	L1 • oui je sais mais euh peut-être que Jean aura une bagnole
ADULTE • ah c'est un Babar	L2 • c'est combien de temps jusqu'à Troyes
VERONIQUE • non c'est un éléphant *[pron=efa~]*	L1 • pff ça fait trois heures je crois c'est long hein
ADULTE • c'est un oui c'est un éphant *{sic}*	L2 mh oh c- ça va

- d'autre part, certaines recherches sur le corpus sont difficiles à formuler parce qu'elles font intervenir des fichiers différents, ce qui est le cas, par exemple, si l'on souhaite trouver tous les corpus qui contiennent une expression donnée (à trouver dans le fichier de transcription) et pour lesquels un des locuteurs est une femme entre 30 et 40 ans (à trouver dans la fiche documentaire).

Ces deux points nous ont poussés vers un format de représentation accumulant dans un même fichier les données et les méta-données.

Un autre problème auquel nous devons faire face est celui de l'hétérogénéité dans les formats initiaux des corpus : fichiers alignés à l'aide de Transcriber vs fichiers transcrits sous Word et donc non alignés.

Tableau 2
Récapitulatif des symboles de transcription

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

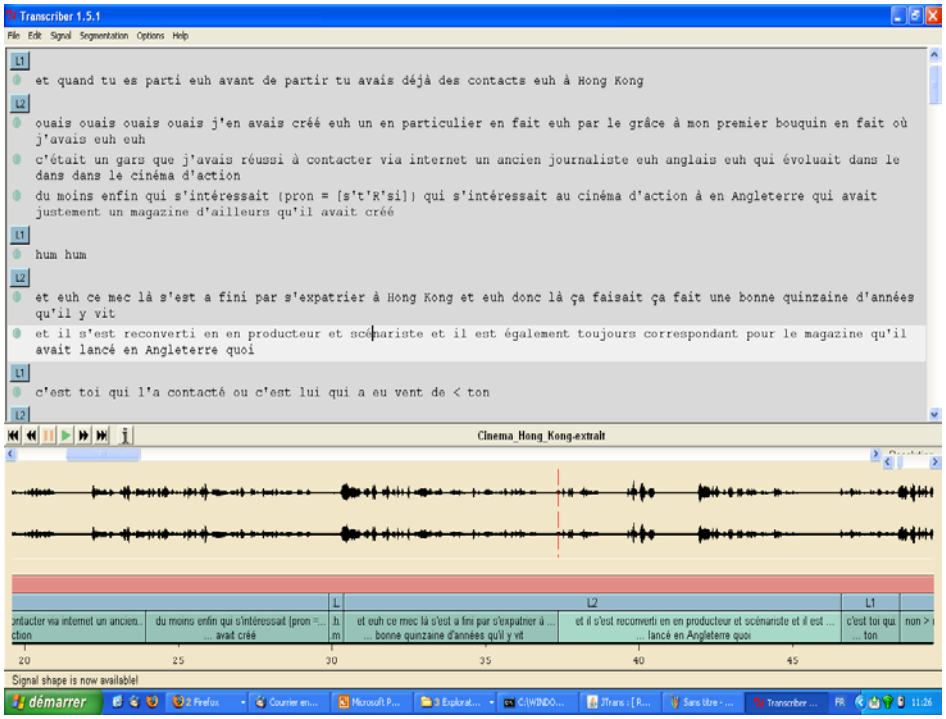

Figure 1. Copie d'écran du logiciel Transcriber

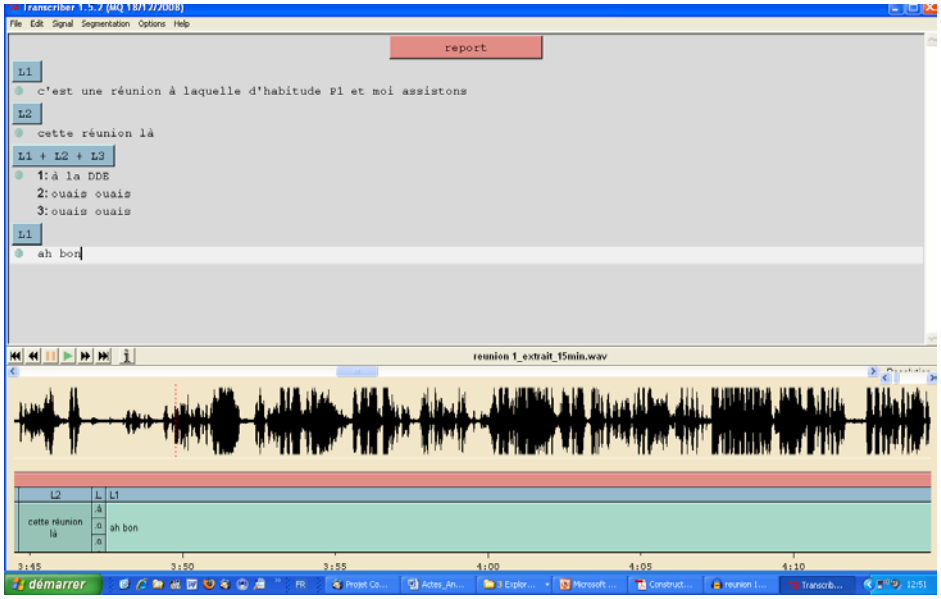

Figure 2. Version modifiée de Transcriber pouvant prendre en compte un nombre non limité de locuteurs dans le cadre de chevauchements de parole (réalisé par Matthieu Quignard)

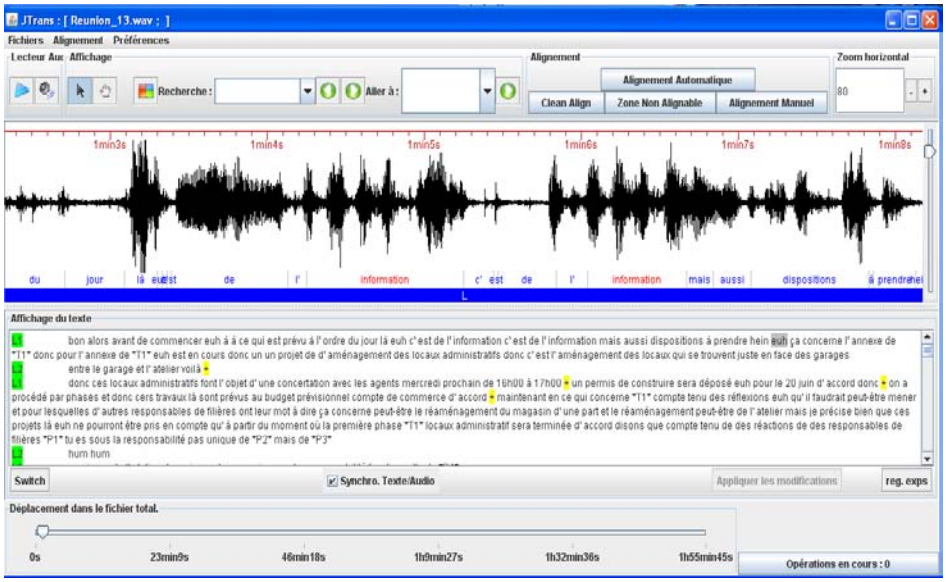

Figure 3. Interface du logiciel JTRANS

Tableau 3
Extrait d'une fiche documentaire

Général	
Identifiant	Anthony2_Can
Ancien identifiant/Titre	
Responsable(s) du corpus	E. Canut
Droit d'accès	Libre
Contact (courriel)	emmanuelle.canut@univ-nancy2.fr
Nom(s) du/des corpus associé(s)	Anthony_Can
Type d'association	Longitudinal
Alignement texte-son	Oui
Nom du logiciel utilisé pour l'alignement	Transcriber
Support(s) anonymisé(s)	Son+transcription
Type d'anonymisation du son	Bip
Notes sur l'anonymisation	
Nombre de locuteurs	2
Relation entre participants	Cadre de l'animation
Type de corpus	adulte-enfant
Modalités de recueil des données	Chercheur participant
Canal de communication	Face à face
Cadre situationnel	Scolaire ou périscolaire
Degré d'interactivité	Très interactif

Ces deux problèmes nous ont poussés à choisir un format de transcription TEI (Text Encoding Initiative[1]) qui permet :

- une intégration des méta-données et des données dans un même fichier;
- un encodage des fichiers alignés texte/son et des fichiers non alignés[2].

Le passage d'un fichier au format Transcriber accompagné de sa fiche documentaire vers un fichier au format TEI est en cours d'automatisation.

Un autre argument pour le passage au format TEI est celui de l'enrichissement du corpus a posteriori, par exemple le cas d'un fichier simplement transcrit que l'on voudrait enrichir par un étiquettage en partie du

discours ou en tours de parole. Dans la mesure où le format TEI est le fruit de plus de vingt-cinq ans de travail de communautés diverses, des modules permettant le type d'annotation que nous venons de mentionner existent. Par ailleurs, le format TEI étant largement documenté (3500 pages de recommandations avec référencement des annotations), l'échange de corpus est facilité.

Il est entendu que la constitution et la normalisation des corpus ne doivent pas nuire à leur lisibilité : pour pouvoir exploiter les corpus à des fins de recherche, la mise en forme doit être facilitée et correspondre aux besoins du chercheur.

L'exploitation linguistique des corpus alignés

Démarche d'exploitation pour les corpus adulte-adulte

Plusieurs démarches sont envisageables selon les objectifs du chercheur. Lorsqu'il s'agit d'analyser la distribution tant lexicale que syntaxique d'une construction, l'équipe a recours à l'analyse des concordances, grâce aux concordanciers existants et libre d'accès sur le net. Elle utilise par ailleurs le logiciel Contextes[3] mis à la disposition de l'équipe depuis 2007.

Le recours aux concordanciers permet de mener des études quantitatives grossières sur la fréquence d'une construction ou les caractéristiques les plus évidentes de ses distributions, et ce en dehors de toute considération sur les types de textes concernés. Le logiciel de recherche Contextes est compatible, moyennant des travaux de programmation supplémentaires, avec la version alignée texte et son du corpus. Il permet donc d'observer les réalisations prosodiques des configurations recherchées. En l'absence de logiciels d'analyse, les observations prosodiques restent sommaires mais peuvent néanmoins mettre en évidence les réalisations particulières qui nécessitent une analyse spécifique. Nous illustrons ce point par un exemple.

Dans un corpus d'environ 350 000 mots (état de notre corpus dans les années 2005), une requête fait apparaître la fréquence d'une réalisation particulière du verbe « *penser* ». Ce verbe est réalisé, dans plus de 90% des cas, sous la forme conjuguée, à la première personne du singulier : « je pense ».

La construction avec « penser » se présente de deux façons :

- une construction en « que phrase » :

  *je **pense** que ça a apporté beaucoup de choses* [athée]

- une construction en « je pense »

  *à peu près toutes les heures il y avait + euh je **pense** euh + un quart d'heure euh ou vingt minutes de récré* [internat] (Voir Figure 4)

Lorsque la construction apparaît entre deux constructions verbales, comme dans l'exemple suivant, il y a deux possibilités de découpage :

*et mais ça n'a rien à voir je **pense** il y a de très bons chefs qui sont des des ouvriers*

[trombone]

On peut en effet regrouper la construction avec le segment gauche :

*mais ça n'a rien à voir **je pense***

Mais, on peut également regrouper la construction avec le segment droit dans l'énoncé :

*je **pense** il y a de très bons chefs qui sont des ouvriers*

C'est l'accès au son qui permet de déterminer à quel segment doit-on rattacher la construction.

Le corpus en cours de constitution n'étant pas étiqueté, la recherche se fait à partir de la stricte chaîne de caractères. Pour faciliter l'analyse quantitative des distributions, le logiciel prévoit un transfert direct vers un tableur Excel qui permet un comptage rapide et un marquage manuel des catégories par le recours à la fonction filtre. La méthodologie permet ainsi de croiser analyse quantitative et qualitative.

Le corpus peut donner lieu également à une exploitation de type exploration. Il s'agit en ce cas de repérer par une lecture exhaustive les phénomènes langagiers recherchés et pour lesquels des facteurs externes peuvent être prépondérants. Dans ce cas, l'exploitation des corpus passe par le choix des interactions à analyser. Elle peut se faire en fonction de la situation de communication dans laquelle se trouvent les interactants. Dans ce cas, le genre de discours (conversation, réunion de travail, transaction commerciale, etc.), le nombre de locuteurs, leur identité ou les relations qu'ils entretiennent, les thèmes abordés ou encore la longueur des échanges peuvent être des critères de choix pour la sélection des corpus à analyser. Cette première étape est très importante car elle a un fort impact sur les résultats dans la mesure où les analyses tiennent compte des conditions de productions des discours oraux.

Par exemple, on constate que les recherches de « dénomination » sont souvent construites à partir d'un schéma récurrent :

euh c'est à dire qu'il y a des
- ce sont pas des immeubles +
- ce sont des petites maisons hein
- c'est en général en auto-construction [Chili]

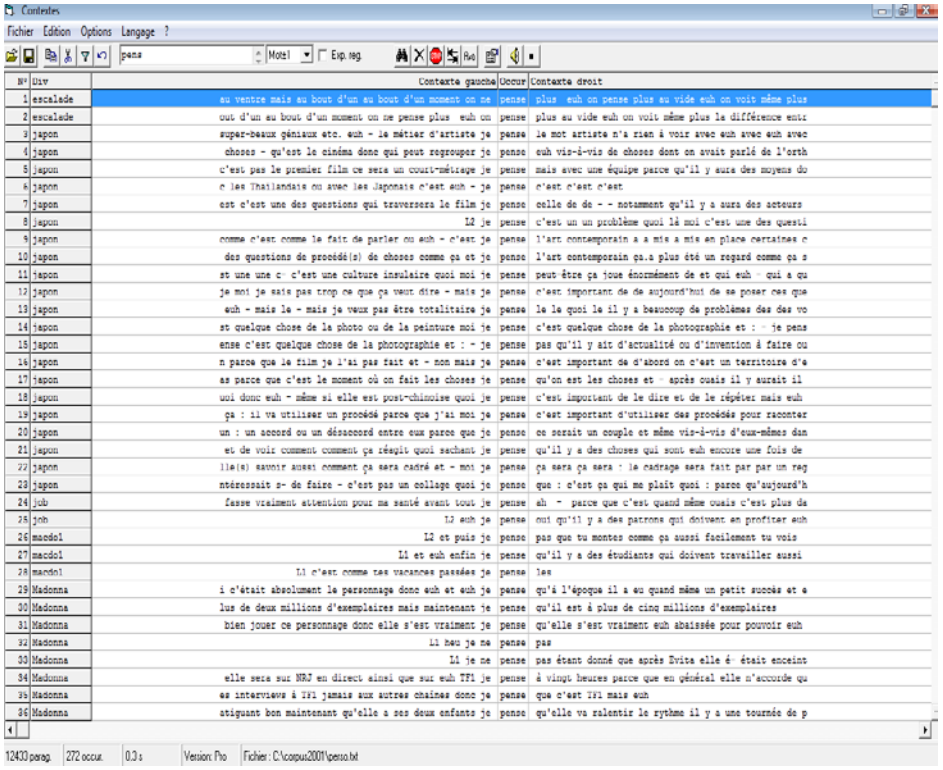

Figure 4. Aperçu des résultats de la requête « penser » avec le logiciel Contextes

il y avait quand même une espèce de
comment dirais-je
- pas des seigneurs mais enfin
- une bourgeoisie qui possédait des terrains
  - qui possédait des maisons
  - qui possédait beaucoup de choses

Il peut être intéressant dans ce cas d'observer l'impact du type de texte concerné dans la constitution de ces définitions.

Ce dernier type d'analyse est possible sur la version papier du corpus à partir d'une exportation en HTML, de la transcription réalisée à l'aide de Transcriber. Une première lecture permet de repérer les phénomènes langagiers intéressants, et d'annoter « à la main » sur la transcription papier. Les corpus

étant alignés texte-son, il est aisé d'atteindre le segment sonore associé au passage choisi.

Les études pouvant être menées dans le cadre de cette approche sont nécessairement limitées à des données relativement restreintes. Il est en effet peu envisageable de traiter « à la main » plusieurs centaines de milliers voire plusieurs millions de mots.

Démarche d'exploitation pour les corpus adulte-enfant

Pour ce qui concerne les corpus d'interaction adulte-enfant, l'analyse recèle quelques spécificités. Nous procédons à une première analyse des énoncés de l'adulte et de l'enfant en grandes catégories syntaxiques : énoncés comportant maximalement une construction « simple », plusieurs constructions « simples » juxtaposées ou coordonnées, une ou plusieurs constructions « complexes ». Le repérage des constructions dites complexes s'effectue en fonction d'une liste d' « introducteurs de complexité » (correspondant globalement à des éléments de subordination). L'annotation des énoncés selon ces catégories est manuelle, soit sur une grille, soit directement sur la version imprimée du corpus (voir Tableaux 4 et 5).

La phase d'analyse quantitative vient ensuite : pour chaque catégorie et dans chaque corpus, nous mentionnons son pourcentage afin de repérer s'il existe une évolution dans le temps (voir Tableau 6).

Une seconde analyse vient alors compléter la première. On repère les constructions complexes pour lesquelles il y a une évolution générale (en termes quantitatifs et qualitatifs). Nous procédons à l'observation non outillée d'éléments qui semblent s'installer dans des contextes différents, mais aussi des reprises par l'enfant des verbalisations de l'adulte (soit sous forme d'essais c'est-à-dire de tentatives non abouties, soit sous forme de reprises immédiates ou différées, identiques ou avec des variations). Par la suite, une attention particulière peut être portée à une construction qui semble pertinente au regard de l'évolution générale en corrélation avec le langage adressé à l'enfant (voir Tableau 7).

Dans cette démarche, c'est essentiellement la phase d'annotation et de repérage des phénomènes qui pourrait être facilitée par des outils informatiques. Pour la phase de repérage, les concordanciers simples (comme Antconc) sont une première aide. Pour la phase d'annotation, des logiciels tels que N'Vivo sont à l'étude actuellement.

Tableau 4
Grille originale des introducteurs de complexité (d'après L. Lentin, 1998)

à + verbe à l'infinitif	17	Ex : elle donne *à manger*
Comme (= étant donné que)	**16**	Ex : *comme* la cloche a sonné, ils sont rentrés dans la classe
Comparative (= comme)	**21**	Ex : elle écrit *comme* elle parle
de + verbe à l'infinitif	18	Ex : elle n'est pas contente *de voir* sa maman
discours indirect (paroles rapportées)	**13**	Ex : elle *dit de* faire la vaisselle ; elle *dit que* je suis beau
divers	**24**	Opposition (tandis que, alors que), tellement que, sans que, surtout que, déjà que, sinon, etc
Extraction	1	C'est, voilà, il y a qui, que, où, etc
Gérondif	**20**	Ex : en marchant
il faut que	4	Ex : il *faut que* j'aille aider mes amis
interrogative indirecte	12	Ex : je sais ce que, je vais voir comment, je (ne) sais (pas) si je vais aller
introducteurs temporels (autres que « quand »)	**7**	dès que, après que, chaque fois que, pendant que, avant que, etc
Où relatif	**14**	Relatif ; ex : elle a trouvé le paquet *où* il y a les couches du bébé
parce que	**8**	Ex : le bébé ne boit pas *parce qu'il* a déjà bu son lait
pour + verbe à l'infinitif	19	Ex : papa sort sa clé *pour ouvrir* la porte
pour que	**15**	Ex : elle appuie un petit peu *pour que* le dentifrice sorte du tube
puisque	**23**	Ex : il va jouer tout seul *puisque* sa copine est partie
Quand (= lorsque)	**6**	Ex : ils sont contents *quand* ça roule
Quantitative (comparatif)	**11**	Plus, moins, autant…que ; ex : Paul est *plus* grand *que* Pierre
Que conjonction	**5**	Ex : elle trouve *que* c'est drôle (« que » après un verbe conjugué)
Que relatif	**22**	Ex : elle ouvre le cadeau *que* son papa lui a donné
Qui relatif	**3**	Ex : il sort du camion un canapé *qui* est très lourd
Si (supposition et condition)	**9**	Ex : *si* le chat griffe le ballon il peut éclater
verbe + verbe à l'infinitif (sauf le futur proche)	10	Ex : il *veut rentrer* dans sa chambre ; il *peut aller* le *chercher*

Tableau 5
Extrait d'une analyse en catégories syntaxique d'énoncés

Prénom et âge de l'enfant : Amel, 6ans 3mois **N° corpus :** 2

Numéro de l'énoncé **E= Enfant** **A= Adulte**	**Énoncés comportant une construction simple**		**Énoncés comportant des constructions simples multiples (juxtaposées ou coordonnées)**		**Énoncés comportant des constructions complexes (cf liste IC)** + : juxtaposition / : emboîtement inc : incomplétude		**Énoncés comportant des essais (cf liste IC)**	**Fiches de commen-taires**
	E	**A**	**E**	**A**	**E**	**A**	**E**	**A/E**
A21				X				
E21			X					
A22						12/10		
E22					5/7/4			Essai 7 ? Pb sémantique (après et non avant)
A23						1/12/10		Aspectuel « finir de »
E23					12 inc			Ou essai ?
A24						19/12 inc		
E24			X					
A25						20/12/13 inc		
E25					20 inc			Essai 20/12 ?
A26						12 inc		
E26			X					
A27		X						
E27							18 pour 5 ou 19	
A28						5		
E28	X							
A29		X						
E29					19			
A30						13		
Total	**11**	**13**	**6**	**3**	**6**	**14**	**6**	

Tableau 6
Évolution de la production des constructions syntaxiques
chez l'adulte et l'enfant

N° du corpus		**1**		**2**	
Adulte (A)		**A**	**E**	**A**	**E**
Enfant (E)		**(/25)**	**(/24)**	**(/30)**	**(/29)**
Constructions simples (%)		64	66,66	43,33	37,93
Constructions simples multiples (%)		4	8,33	10	20,69
Constructions	**Total**	32	25	46,66	41,38
complexes (%)	*Réalisées*	*32*	*16,66*	*46,66*	*10,34*
	En incomplétude	*0*	*0*	*20*	*6,9*
	En essais	*-*	*8,33*	*-*	*20,68*
	Comportant 2 IC	*0*	*0*	*10*	*3,44*
	Comportant plus de 2 IC	*4*	*0*	*6,66*	*3,44*

Tableau 7
Exemple de début d'analyse tirée des tableaux précédents

Commentaires sur le langage de l'enfant :

- Augmentation du pourcentage de constructions simples multiples et de constructions complexes (essais compris).

- Pour les constructions syntaxiques complexes : forte augmentation des essais (l'enfant tente plus de constructions) et dans le même temps, apparition des constructions comportant plus d'un introducteur de complexité. Phénomènes à observer : les tentatives sur « pendant que ».

Commentaires sur le langage de l'adulte :

- Evolution en parallèle de l'enfant : augmentation du pourcentage de constructions simples multiples et de constructions complexes.

- Hypothèses : les propositions de l'adulte se situent dans la zone proche du développement syntaxique de l'enfant et l'amène ainsi à tenter des constructions non encore maîtrisées comme les subordonnées temporelles : faire l'analyse qualitative des interactions (reprises et reformulations).

Prospectives

Nous aimerions en guise de conclusion poser quelques questions sur le potentiel des logiciels :

- faut-il un outil idéal? Pour faire quoi? Un logiciel qui remplacerait complètement l'annotation manuelle est-il envisageable? Souhaitable?

Les chercheurs interviennent toujours manuellement sur leur corpus car il paraît peu probable, dans l'état actuel des connaissances, que par exemple le repérage des questions ou encore des recherches de dénomination soit réalisable;

- n'est-ce pas illusoire d'attendre un outil générique adapté à tous les utilisateurs? Ne faut-il pas plutôt des outils plus performants qui serviraient à l'ensemble de la communauté travaillant aussi bien sur des données orales qu'écrites (un logiciel de type Contextes libre et gratuit)?
- cette question des logiciels est-elle essentielle? En effet, les chercheurs n'utilisent souvent qu'une partie ou détournent les fonctionnalités des logiciels pour leur usage propre et il en sera probablement toujours ainsi. N'est-ce pas plutôt la question du cumul des connaissances par le cumul des exploitations qui est à envisager comme priorité?

Notes

[1] Les formats Transcriber et TEI sont l'un et l'autre définis comme des vocabulaires XML. Sur le plan pratique, cela signifie qu'on passe de l'un à l'autre via des outils classiques dans le monde XML comme des feuilles de style XSL.
[2] Dans un fichier non aligné texte-son, on peut néanmoins savoir que la prise de parole d'un second interlocuteur débute après un mot précis émis par le premier interlocuteur; le format TEI permet également la prise en compte de ce type d'information.
[3] Logiciel créé par Jean Véronis (Université de Provence). Nous le remercions.

Références

Baude, O. (Éd.). (2006). Corpus oraux. Guide des bonnes pratiques. Paris : CNRS Éditions.

Benzitoun, C., Campione, E., Deulofeu, J., Henry, S., Teston, S., Valli, A., & Véronis, J. (2004, 15 Mai). *L'analyse syntaxique de l'oral : problèmes et méthode.* Communication présentée aux Journées d'Études de l'ATALA, Evans : Méthodes et outils pour l'évaluation des analyseurs syntaxiques, Paris.

Bilger, M. (Éd.). (2000). *Linguistique sur corpus. Études et réflexions.* Perpignan, France : Presses universitaires de Perpignan.

Canut, E. (Éd.). (2008). Des lectures partagées aux recherches sur corpus. Aperçu de réflexions actuelles. *L'Acquisition du langage oral et écrit* (ALOE), [Numéro thématique], *60-61*.

Cappeau, P., & Seijido, M. (2005). *Les corpus oraux en français (inventaire 2005 v.1.0)*. Document consulté le 20 septembre 2008 de http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire. Pdf.

Debaisieux, J.-M. (2005). Les corpus oraux : situation, exploitation linguistique. Bilan et perspectives. *Scolia, 19*, 9-40.

Équipe DELIC (2004). Autour du Corpus de référence du français parlé. *Recherches sur le français parlé, 18*, 11-42.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus.* Paris : Colin.

Savelli, M. (Éd.). (2005). Corpus oraux et diversité des approches. *Lidil,* [Numéro thématique], *31*.

Véronis, J. (2000). Annotation automatique de corpus : panorama et état de la technique. Dans J.M. Pierrel (Éd.), *Ingénierie des langues* (pp. 111-129). Paris : Hermès.

Références complémentaires

Liste des sites internet des institutions, bases de données, corpus et logiciels :

Analyse et traitement informatique de la langue française (ATILF), UMR CNRS : www.atilf.fr
Base textuelle FRANTEXT : http://atilf.atilf.fr/frantext.htm
British national corpus : www.natcorp.ox.ac.uk
Centre de ressources pour la description de l'oral (CRDO) : crdo.risc.cnrs.fr/
Centre national de ressources textuelles et lexicales : http://www.cnrtl.fr/
Corpus de langue parlée en interaction (CLAPI) : http://clapi.univ-lyon2.fr
Corpus de référence de l'espagnol actuel (CREA) : pheme.rae.es/java.ext/corpus.htm
Corpus search, management and analysis system (COSMAS II) : http://www.ids-mannheim.de/ cosmas2/
Délégation générale à la langue française et aux langues de France : http://www.culture.gouv.fr/culture/dglf/
Enquête sociolinguistique à Orléans (ESLO) : http://www.univ-orleans.fr/eslo/
Logiciel ANTCONC : http://www.antlab.sci.waseda.ac.jp/antconc_index.html
Logiciel JTRANS : http://jtrans.gforge.inria.fr
Logiciel TRANSCRIBER : http://trans.sourceforge.net/
Projet PFC (Phonologie du français contemporain) : http://www.projet-pfc.net/

Virginie André *est maître de conférences à l'Université Nancy 2 et rattachée au laboratoire CNRS ATILF. Docteur en Sciences du langage, ses travaux portent plus particulièrement sur l'analyse des interactions verbales et la construction collaborative du discours notamment en situation de travail. Ses domaines d'intérêt concernent également la constitution, l'exploitation et la diffusion des corpus oraux ainsi que l'implication et l'application des sciences du langage, notamment dans le domaine de la formation linguistique.*

Christophe Benzitoun *est maître de conférences à l'Université Nancy 2 et rattaché au laboratoire CNRS ATILF. Docteur en Sciences du langage, diplômé de l'Université de Provence, il travaille plus particulièrement sur l'analyse syntaxique du français parlé et la question de l'informatisation des données orales.*

Emmanuelle Canut *est maître de conférences à l'Université Nancy 2 et rattachée au laboratoire CNRS ATILF. Docteur en Sciences du langage, elle s'intéresse à l'acquisition du langage par l'enfant et notamment aux interactions entre adulte et enfant dans la mise en fonctionnement du langage. Cette recherche a des implications dans le domaine scolaire et périscolaire avec la mise en œuvre d'actions et de formations à destination des professionnels de l'enfance.*

Jeanne-Marie Debaisieux *est maître de conférences au département de Sciences du langage de l'Université Nancy 2 et rattachée au laboratoire CNRS ATILF. Ses travaux portent sur l'analyse syntaxique et pragmatique de la langue parlée et sur la méthodologie de recueil et de traitement des données orales, tant à des fins de recherches linguistiques que d'applications en didactique des langues.*

Bertrand Gaiffe *est chargé de recherche dans l'UMR ATILF rattachée à l'Université de Nancy 2. Docteur en informatique, il travaille plus particulièrement la constitution, l'annotation, la normalisation, le versionnage de données textuelles de nature variée. Dans un cadre plus large, il est fortement impliqué dans le projet européen CLARIN dont l'un des objectifs est de fournir des données textuelles et sonores et des plateformes de traitements à l'échelle européenne.*

Evelyne Jacquey *est chargée de recherche dans l'UMR ATILF rattachée à l'Université de Nancy 2. Docteur en linguistique-informatique, elle travaille plus particulièrement sur l'analyse sémantique du français, écrit et parlé, par le biais de traitements et d'annotations de corpus. Elle s'intéresse aussi à la constitution de corpus écrits et oraux sur le français.*