Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL)

Michel Bert, Sylvie Bruxelles, Carole Etienne, Emilie Jouin-Chardon, Justine Lascar, Lorenza Mondada, Sandra Teston et Véronique Traverso

Édition électronique

URL : <http://journals.openedition.org/pratiques/1581>

DOI : 10.4000/pratiques.1581

ISSN : 2425-2042

Éditeur

Centre de recherche sur les médiations (CREM)

Édition imprimée

Date de publication : 15 décembre 2010

Pagination : 17-34

Référence électronique

Michel Bert, Sylvie Bruxelles, Carole Etienne, Emilie Jouin-Chardon, Justine Lascar, Lorenza Mondada, Sandra Teston et Véronique Traverso, « Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL) », *Pratiques* [En ligne], 147-148 | 2010, mis en ligne le 15 décembre 2010, consulté le 20 avril 2019. URL : <http://journals.openedition.org/pratiques/1581> ; DOI : 10.4000/pratiques.1581

Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL)

Groupe ICOR

Michel Bert, Sylvie Bruxelles, Carole Etienne,
Emilie Jouin-Chardon, Justine Lascar, Lorenza Mondada,
Sandra Teston, Véronique Traverso

Laboratoire ICAR
Université de Lyon & CNRS

1. Introduction

Le développement de banques de données consacrées à l'oral⁽¹⁾, observé internationalement depuis les années 1990, et auquel la France participe, non sans quelque retard, est un des indices du rôle fondamental qu'occupent désormais les corpus à la fois comme outils de recherche pour la communauté des linguistes et comme ressources pour une variété de perspectives et de disciplines en sciences humaines et sociales.

Cet article se penche sur un type particulier de corpus, constitué d'enregistrements audio et/ou vidéo d'interactions en contexte social authentique. Ce type de corpus fonde la linguistique interactionnelle, mais est de plus en plus recherché par d'autres domaines de la linguistique, comme, par exemple, la syntaxe fonctionnelle ou l'enseignement du français langue seconde, attentifs à décrire la langue telle qu'elle est véritablement pratiquée en contexte. Ces corpus sont constitués de données recueillies en situation « naturelle », c'est-à-dire dans le contexte social ordinaire des activités documentées, en dehors de toute élicitation ou expérimentation. Ces corpus naturalistes ou écologiques d'interactions

(1) Pour un bilan historique, cf. Bruxelles, Mondada, Simon, Traverso (2009).

sont aujourd'hui archivés, mutualisés et valorisés dans des banques de données multimédia, accessibles via Internet, qui ont proposé des solutions aux problèmes technologiques et théoriques spécifiques à ce type de données.

Cet article se propose de contribuer à la réflexion sur les enjeux théoriques et analytiques liés aux entreprises de constitution des banques de corpus de parole en interaction. Il se fonde sur les expériences menées au laboratoire ICAR qui ont conduit à la mise en ligne de la plateforme Corpus de Langue Parlée en Interaction (CLAPI⁽²⁾) et d'une base de connaissances et de ressources qui lui est associée, CORINTE⁽³⁾. Il fait également référence aux travaux actuellement menés sur le Corpus International Ecologique de la Langue Française (CIEL-F⁽⁴⁾), en cours de constitution dans le cadre d'un projet franco-allemand ANR-DFG portant sur la comparaison des variétés de Français Parlé en Interaction (FPI) dans le monde francophone.

En décrivant la manière dont sont conçus ces grands corpus, l'article parcourt d'abord un certain nombre de questions posées par l'archivage des corpus dans une base de données consultable en ligne, questions relatives tant à la constitution des données primaires et secondaires, qu'à leur documentation et à leur accessibilité soit directe soit à travers l'interconnexion avec d'autres bases. Dans une deuxième partie, on envisage les options actuellement développées dans CLAPI pour la mise à disposition d'outils d'interrogation destinés au traitement de grandes masses de données d'une part, et d'autre part spécifiquement dédiés à l'analyse de données interactionnelles.

2. Recueil et constitution de grands corpus de Langue Parlée en Interaction (LPI) : les exemples de la plateforme CLAPI et du projet CIEL

Dans cet article, nous nous basons sur les expériences réalisées par la banque de données CLAPI depuis plus d'une décennie et dans le cadre du projet CIEL, plus récent. Avant de présenter un certain nombre de problématiques qu'il faut résoudre quand on s'attelle au travail important de constitution d'une banque de données de corpus de parole en interaction, nous présentons brièvement ces deux initiatives⁽⁵⁾.

(2) <http://clapi.univ-lyon2.fr/>.

(3) <http://icar.univ-lyon2.fr/projets/corinte/>. Ce site explicite et conceptualise toutes les étapes de la constitution et de l'exploitation des corpus. Il expose également la méthodologie et les arrière-plans théoriques de la démarche du travail sur corpus dans une perspective interactionnelle, et il propose diverses ressources (bibliographies, informations sur les logiciels d'alignement, la compression ou les formats audio/vidéo), ainsi que des entrées thématiques, notamment sur les questions juridiques.

(4) Le projet CIEL-F est le fruit d'un partenariat entre le laboratoire ICAR et S. Pfänder à Freiburg, A.C. Simon à Louvain, F. Gadet à Paris/Modyco, R. Ludwig à Halle. <http://www.ciel-f.org/>.

(5) Pour des initiatives similaires, voir la *TalkBank* (responsable B. McWhinney), le *Santa Barbara Corpus of Spoken American English* (responsable J.W. Du Bois), le *MOVIN Corpus* et le futur corpus de référence danois (responsable J. Wagner) ainsi que le corpus FOLK de l'Institut für Deutsche Sprache Corpus (responsables A. Depperman & M. Hartung).

2.1. La plateforme CLAPI

Le projet CLAPI a démarré à la fin des années 1990 au sein de l'UMR GRIC (Groupe de Recherches sur les Interactions Communicatives) de Lyon dans le but d'inventorier et de rendre consultable la masse de corpus confectionnés par les chercheurs et les étudiants de cette unité, devenue aujourd'hui l'UMR ICAR (Interactions, Corpus, Apprentissages, Représentations). Une première base de données, CLAPI 1.0 a été mise en ligne en 2000, recourant à une trentaine de champs permettant de décrire les corpus, accompagnés d'extraits courts d'enregistrements et de leurs transcriptions (pour une description de l'état du projet en 2001, voir Bruxelles & Traverso 2002). Grâce à plusieurs programmes de recherche, le projet CLAPI a ensuite connu un nouveau développement (voir Balthasar & Bert 2005) permettant le lancement en 2005 de la plateforme CLAPI 2.0. Elle comporte l'intégralité des corpus de la base précédente, numérisés et décrits avec davantage de métadonnées. Elle inclut des corpus fondateurs pour l'étude de l'interaction en France, comme *Mode*⁽⁶⁾ et *La Dame de Caluire*⁽⁷⁾, et les différents corpus élaborés depuis cette date à ICAR et dans d'autres laboratoires et institutions, pour y être mis à la disposition de la communauté des chercheurs ; c'est le cas par exemple du *Fonds Bielefeld*⁽⁸⁾, un autre corpus historique pour l'étude de l'interaction exolingue entre natifs et non-natifs. CLAPI offre actuellement un hébergement à tout corpus de parole en interaction présentant un intérêt pour la communauté scientifique.

Un ensemble de 600 heures de données interactionnelles a ainsi été archivé, dont 135 heures ont été intégrées à la base de données⁽⁹⁾. Cet ensemble est particulièrement riche du fait de la diversité des situations sociales documentées : on y trouve des situations classiquement étudiées en linguistique interactionnelle, comme les conversations quotidiennes et les interactions de travail, mais aussi des situations qui figurent très rarement dans les banques de données (par exemple, soins dentaires, négociations des biens au cours de divorces), voire des situations qui couvrent une grande diversité de pratiques sociales (comme les interactions de service, les échanges dans des commerces, etc.). La valeur de ces corpus réside aussi dans la richesse des données primaires (comprenant outre des enregistrements audio, un nombre croissant de données vidéo, voire de données multiscopes produites grâce à des technologies sophistiquées assurant le montage synchronisé de sources audio et vidéo multiples). Une partie de ces données, soit 58 heures, correspondant à 125 transcriptions alignées avec le signal, ont été formatées en XML pour être requêtables⁽¹⁰⁾.

La deuxième innovation de CLAPI 2.0 – qui transforme la base de données en une véritable plateforme « outillée » – est d'avoir ajouté à la fonction d'archivage d'origine une visée d'exploitation : rendre possible l'interrogation de la

(6) Voir, pour des analyses croisées de ce corpus, Cosnier & Kerbrat-Orecchioni (1986).

(7) Voir, pour des analyses croisées de ce corpus, Bange (1987).

(8) Ensemble de corpus collectés par E. Gülich, U. Dausendschön-Gay et U. Kraft de l'université de Bielefeld, consultables sur CLAPI.

(9) Soit 300 enregistrements et 400 transcriptions, chiffres en constante évolution.

(10) Le format XML a aussi permis de gérer l'hétérogénéité des transcriptions et les différentes conventions de transcription utilisées, en permettant d'identifier de manière homogène un certain nombre de phénomènes sur lesquels fonder les requêtes.

base de données grâce à des outils de requêtes et permettre ainsi une approche analytique des corpus, dont nous exposons les grandes lignes dans la section 4.

Actuellement, l'équipe est en train de développer une version 3.0 de CLAPI qui vise à simplifier l'alimentation et la mise à jour de la banque de données, à la rendre capable d'accueillir des données aussi complexes que les corpus plurilingues et les corpus multimodaux annotés ; elle offrira une meilleure ergonomie, et de nouveaux outils avec la prise en compte des profils d'utilisateurs afin d'améliorer la pertinence des requêtes. Ces développements tiennent compte de l'analyse des traces des consultations et des requêtes (entre 10 000 et 20 000 accès chaque mois) et des retours exprimés directement par les utilisateurs (option « Contact » présente dans chaque menu).

2.2. Le projet CIEL

Le projet CIEL entend également mettre à la disposition des chercheurs des données écologiques, mais il comporte des spécificités qui en font une entreprise particulièrement innovante. La collecte des données « naturelles » s'est organisée autour d'un objectif spécifique : comparer la diversité des français parlés dans le monde francophone sur la base de situations d'interaction ciblées afin de permettre des analyses comparatives. Trois situations ont été retenues pour la collecte : les conversations à table ; les réunions de travail ou interactions de service/commerciales ; les enregistrements d'appels ou de débats à des radios locales. Ces activités seront documentées dans des aires communicatives distribuées à travers le monde, notamment en France, Suisse, Belgique, Antilles, Canada, Algérie, Egypte, Côte d'Ivoire, Cameroun, Sénégal, Liban, Ile Maurice, Ile de la Réunion, grâce à des collaborations avec un réseau de chercheurs locaux. Le croisement d'un principe « praxéologique », définissant les corpus à recueillir en termes de types d'activité, et d'un principe d'« aires » pour définir les espaces communicatifs à couvrir (Dister *et alii*, 2008 ; Gadet, Ludwig, Pfänder, 2008) définit le corpus CIEL-F de manière originale par rapport à d'autres corpus, dont la structuration se base sur des critères de représentativité sociolinguistique (concernant soit des types de locuteurs soit des types de discours). Le projet se construit sur une double expertise : d'une part il s'appuie sur une expertise internationale en matière de bases de corpus – fondée sur la banque de données CLAPI, la banque de données Valibel développée à l'université Louvain-La-Neuve et le système d'exploitation MOCA, développé conjointement par Freiburg et Louvain. D'autre part, il s'appuie sur les expertises différenciées des partenaires dans les domaines de la linguistique interactionnelle (laboratoire ICAR), la sociolinguistique (laboratoire MoDyCo, Centre Valibel) et la linguistique de contact (Groupe contact de cultures, Halle Universität ; Romanisches Seminar, Freiburg Universität). Les corpus du projet CIEL sont actuellement en voie de constitution et de standardisation afin d'être intégrés dans la banque de données ; ils seront accessibles à la fin du projet.

Si la plateforme CLAPI s'est développée d'abord comme un système d'archivage de corpus collectés par ailleurs, puis s'est transformée en un système d'exploitation outillé, le projet CIEL est centré sur la constitution d'un corpus cohérent et a prévu dès le départ un protocole de collecte de données garantissant un recueil aussi homogène que possible dans des aires très différentes. La première

approche a permis à CLAPI de s'enrichir de corpus extrêmement divers – ce qui aurait été impossible dans le cadre d'un recueil gouverné par des typologies pré-définies de corpus représentatifs –, la seconde permet d'enrichir la base existante avec des recueils plus systématiques. Les deux entreprises rencontrent, à différents niveaux, des enjeux de standardisation des données et des procédures.

3. Les savoir-faire requis par la constitution d'une banque de données de LPI

Les corpus hébergés dans CLAPI, comme ceux qui sont actuellement recueillis dans le cadre du projet CIEL, sont conçus en tenant compte des exigences fortes développées par les approches interactionnelles en termes de naturalité. Ils résultent d'une démarche ethnographique et leur collecte repose donc sur un travail de terrain ⁽¹¹⁾ sur lequel nous ne reviendrons pas dans cet article (mais voir Détienne & Traverso, 2009 ; Mondada, 2005, 2006 pour une discussion de ces aspects). Nous allons ici nous concentrer sur les questions juridiques et techniques spécifiquement posées par les banques de données de grands corpus de parole en interaction.

3.1. La composition des corpus

Un corpus est un ensemble d'enregistrements, réunis selon des critères définis par leur auteur (par exemple, le fait d'avoir été enregistrés sur le même site, ou bien de documenter un même type d'activité dans des sites différents), qui articulent des données primaires et des données secondaires.

Les données primaires sont les enregistrements de l'événement. Pour un événement donné, une multiplicité de données primaires est possible : l'événement peut être enregistré par une seule source ou par des sources multiples, audio et/ou vidéo. D'autres éléments viennent nourrir les données primaires, comme les objets manipulés durant l'enregistrement (documents et visualisations pré-existants, créés ou modifiés au cours de l'interaction) ; peuvent également être prélevées sur le terrain les traces informatiques dans le cas d'usages des technologies (par enregistrement dynamique d'écran d'ordinateur). L'alignement de ces artefacts et de l'interaction en fait non plus des objets figés, statiques, mais des objets vivants et « temporalisés » en action (Détienne & Traverso, 2009 ; Mondada, 2007).

Ces différentes sources peuvent être archivées sous différents formats, plus ou moins compressées, dans leur forme brute ou sous une forme éditée (multiscopie), ce qui multiplie ultérieurement les données primaires.

Les données secondaires sont les transcriptions et les descriptions des données. Elles aussi sont multiples : différentes versions de la transcription peuvent être effectuées, transcriptions de base ou avancées. Elles sont associées aux con-

(11) Les principales étapes de ce travail de terrain sont exposées sur le site Corinte : <http://icar.univ-lyon2.fr/projets/corinte/recueil/entree.htm>.

ventions de transcription et aux métadonnées, qui décrivent à la fois les conditions de recueil et les caractéristiques de l'événement documenté.

Dans une banque de données, il est essentiel que tous ces documents soient structurés de manière à fournir une image aussi complète que possible du corpus.

3.2. Aspects juridiques de l'archivage et de la diffusion des données primaires & secondaires

Les enjeux juridiques traversent tous les paliers de la constitution du corpus⁽¹²⁾. Dès les données primaires, ils se posent lors de l'accès au terrain, quand il s'agit de recueillir le consentement éclairé des participants ; lors des enregistrements, où ils sont liés au respect de la vie privée, de la volonté des participants et du choix conséquent de ce qu'il s'agit de montrer ; lorsqu'il s'agit d'anonymiser les enregistrements sonores ou visuels, en insérant des *beeps* ou par des techniques de floutage de l'image.

Les questions juridiques se posent aussi à propos des données secondaires : d'une part, lors de la phase de transcription, notamment sur les modalités d'anonymisation des participants et de toute information jugée comme pouvant porter atteinte à leur vie privée ; elles concernent par exemple le choix de pseudonymes appropriés, qui empêchent d'identifier la personne tout en gardant certaines informations (comme les connotations ethniques, la longueur et certaines caractéristiques syllabiques du nom). D'autre part, les questions juridiques concernent les métadonnées, obligatoirement standardisées dans une perspective documentaire d'archivage et de diffusion.

3.2.1. Contraintes sur la documentation des données

Le choix des métadonnées doit répondre à plusieurs impératifs : décrire précisément les corpus et assurer l'efficacité des requêtes ; faire un compromis entre une standardisation permettant le développement cumulatif de la base et une flexibilité indispensable pour rendre compte des spécificités de chaque corpus. Les métadonnées retenues⁽¹³⁾ – compatibles avec les standards internationaux TEI, Dublin Core/Olac, IMDI – présentent le corpus et documentent les opérations successives de recueil, confection et analyse des données : modalités et formats d'enregistrement, caractéristiques des interactions, propriétés et accessibilités des données primaires et secondaires, bibliographie liée au corpus.

Pour des raisons pratiques, éthiques ou juridiques, certains champs peuvent ne pas être remplis, contenir des informations évasives (région plutôt que localisation précise par exemple) ou ne pas être affichés pour un utilisateur extérieur.

3.2.2. Contraintes juridiques et accès aux données

Les exigences éthiques ou juridiques, relevant du respect de la vie privée, de

(12) Voir la rubrique « Questions juridiques » du site CORINTE, et également Baude (2006).

(13) Dans CLAPI une grille de 75 descripteurs a été élaborée, compatibles avec les 50 prévus dans CIEL.

la propriété intellectuelle ou de la réglementation relative aux bases de données, configurent aussi l'interrogation d'une base puisqu'elles contraignent le caractère diffusable et accessible des corpus. Elles doivent être prises en compte lors de la définition des conditions de dépôt ou legs de corpus aux banques de données, ou lors de la définition de profils d'utilisateurs.

Le dispositif de Clapi donne un accès libre aux métadonnées, et soumet l'accès aux données originelles, primaires ou secondaires, à un profil approprié tant au demandeur qu'aux corpus concernés :

- le profil *Invité* désigne tout utilisateur qui accède à la plateforme sans authentification préalable. Il peut connaître l'état du lexique et sa fréquence d'utilisation pour les transcriptions balisées en XML, télécharger cinquante enregistrements (soit 17 h de signal audio ou vidéo et 70 transcriptions), et surtout bénéficier de la totalité des outils d'interrogation sur 70% de la base.
- toute personne ayant signé une convention d'échange consulte CLAPI sous le profil *Contractuel*. Ces usagers peuvent accéder aux transcriptions et aux enregistrements mentionnés dans la convention aux conditions définies par celle-ci ;
- le profil *Responsable de corpus* désigne les dépositaires de corpus. Ils ont accès à la totalité de leurs données et à toutes les fonctionnalités des outils d'interrogation.

En parallèle de la base de données accessible en ligne, un *espace de travail privé* est proposé aux chercheurs pour leur permettre de bénéficier des mêmes outils de requête, avec un jeu restreint de descripteurs, sur des corpus qui ne sont pas visibles sur CLAPI car en cours de traitement ou de transcription. A la fin de ces travaux, le chercheur peut choisir de basculer ces données de l'espace de travail à la base CLAPI.

3.3. Aspects techniques du traitement et de la diffusion des corpus

Nous insisterons ici sur les transcriptions et sur les enjeux de la standardisation et de l'interopérabilité pour l'exploitation des corpus de LPI.

3.3.1. Alignement des données primaires et de la transcription

Nous avons déjà souligné que les corpus actuels se caractérisent par la multiplicité de sources primaires. Les spécificités de l'analyse de l'oral en interaction font que l'attention se porte non seulement sur les formes de la langue orale (notamment le caractère co-construit des productions verbales : alternance des tours de parole, chevauchements, temporalité des productions, etc.), mais également sur la dimension située, incarnée et matérielle de ces interactions, en tant qu'activités constitutives d'un processus praxéologique ancré dans son contexte et son écologie spécifiques. Cette conception incarnée (*embodied*) de la parole en interaction a plusieurs conséquences. Il en résulte qu'il est essentiel pour l'analyse de disposer de la multimodalité grâce à l'accès au signal vidéo – et pas seulement audio. C'est pourquoi les corpus de CLAPI sont alignés et offrent une synchronisation du signal et de la transcription, qui est réalisée soit grâce à un lo-

giciel d'alignement, comme CLAN ou ELAN, soit générée automatiquement par calcul sur la base de repères temporels réguliers contenus dans les transcriptions. Les occurrences identifiées par les moteurs de recherche sont présentées dans une interface qui permet aussi de les afficher de manière alignée.

3.3.2. Le balisage des transcriptions dans CLAPI

Les contraintes technologiques sont déterminantes pour ce qui est de l'exploitation des données secondaires. Nous ne pouvons discuter ici des principes régissant les transcriptions en linguistique interactionnelle (voir Détienne & Traverso, 2009, Mondada, 2008) et nous nous focalisons plutôt sur le traitement informatique des transcriptions.

CLAPI comporte deux types de corpus : des corpus nouveaux transcrits selon la convention ICOR, et des corpus anciens, qui ont été récupérés auprès de divers chercheurs, ayant travaillé à des époques et dans des paradigmes divers. Ces derniers présentent une hétérogénéité, susceptible de rendre problématique leur exploitation informatique, liée à l'usage de l'orthographe adaptée, ainsi qu'à la diversité des conventions utilisées dans les transcriptions.

D'une part, l'usage de l'orthographe adaptée pour noter des spécificités de la langue orale est très fréquent dans les traditions interactionnistes ; il est souvent décrié en linguistique de corpus en invoquant le fait qu'il rendrait impossibles les recherches automatiques de suites de caractères standards. Dans CLAPI, la reconnaissance de l'identité entre un *token* et sa forme adaptée (*bonjour* et *b'jour*) est assurée par le développement d'une application automatique de reconnaissance des variantes graphiques d'un *token*, fondée sur la technique des n-grammes, qui rapproche ainsi *'tends/'ttends* de *attends*, *'fin* de *enfin*, etc. Les formes non reconnues (*chais* pour *je sais*) sont stockées dans un glossaire.

D'autre part, CLAPI accueille des transcriptions dont les conventions sont l'expression de choix opérés par différents chercheurs dans le cadre d'approches ou d'objectifs spécifiques. Nombre d'entre elles ont également fortement évolué au cours des années (la base a une profondeur historique d'environ 30 ans). En conséquence, les transcriptions sont relativement disparates. Pour permettre des requêtes automatiques sur un matériau de fait hétérogène, l'intégration de toute nouvelle transcription dans CLAPI est réalisée par l'équipe médiathèque grâce à un processus semi-automatique basé sur l'analyse de la convention, la vérification automatique des formes orales détectées, et un jeu de paramètres permettant d'évaluer le niveau de granularité. Cette procédure ne peut être entièrement automatisée et demande une validation par des experts pour garantir une qualité des données indispensable à l'exploitation outillée.

Les transcriptions récentes réalisées par le groupe ICOR sont transcrites selon la convention du même nom. Cette convention ⁽¹⁴⁾ a été développée sur la base des conventions internationalement retenues en linguistique interactionnelle (cf. Jefferson 2004 pour l'analyse conversationnelle, ainsi que Selting *et alii* 1998 pour une variante développée en Allemagne, la convention GAT). Elle vise

(14) Convention disponible sur le site CORINTE : http://icar.univ-lyon2.fr/projets/corinte/bandeau_droit/convention_icor.htm

une transcription de base solide représentant les phénomènes temporels fondamentaux de la parole en interaction (chevauchements, pauses, hésitations, troncations, et autres détails segmentaux). Elle est implémentée en XML. La standardisation de la transcription de la multimodalité est actuellement à l'étude.

Le projet CIEL répond à ces exigences de manière différente : contrairement à CLAPI qui archive des données anciennes, hétérogènes, complétées par des corpus récents transcrits de manière homogène, CIEL est constitué de corpus récoltés en adoptant un même protocole, transcrits dans une convention spécifique proche de la convention ICOR, alignés à l'aide du logiciel Praat, et prévoyant une série de *tiers* pour l'orthographe standard, l'orthographe adaptée, les traductions et les gloses grammaticales (la transcription multimodale n'étant pas prévue). Les sources d'hétérogénéités dans ce dernier cas viennent plutôt des différences de pratiques de notation des différentes équipes sur le terrain et des problèmes soulevés par l'extrême variation des données (variétés régionales et de contact du français dans le monde). En amont des différences de conception entre les deux projets, CLAPI et CIEL montrent l'inéluctable indexicalité du traitement des données, même lorsque des protocoles standards sont adoptés.

3.3.3. Compatibilité et interopérabilité avec d'autres bases de données

En absence de grand projet unifié pour une banque de données de référence en France, et en présence de nombreuses initiatives, certaines récentes, d'autres consolidées, en matière de banques de données de corpus en France et sur la scène internationale, l'enjeu de l'interopérabilité est crucial pour désenclaver les initiatives existantes et éviter l'émiettement des corpus dans des banques de données limitées, locales et isolées.

La plateforme CLAPI travaille actuellement avec les plateformes analogues qui, dans le paysage international, sont spécialisées dans le traitement des corpus de LPI, notamment la TalkBank⁽¹⁵⁾ et les projets de corpus nationaux en Allemagne et au Danemark. Elle travaille aussi en collaboration avec les banques de données qui offrent des corpus en français accessibles sur Internet sous la forme d'architectures robustes et stables, comme ESLO⁽¹⁶⁾, PFC⁽¹⁷⁾, Valibel⁽¹⁸⁾. Le projet CIEL renforce cette recherche d'interopérabilité, puisqu'il s'appuie sur deux banques de données, CLAPI et Valibel, et deux systèmes d'exploitation, CLAPI et MOCA.

Les enjeux d'interopérabilité posent de façon cruciale la question de la standardisation des différentes composantes des corpus, ainsi que de la convertibilité des formats – concernant notamment les données secondaires, métadonnées et transcriptions.

En ce qui concerne les métadonnées, CLAPI a développé une grille propre de descripteurs, adaptés aux exigences de la linguistique interactionnelle. Cependant, pour garantir les échanges avec les autres banques de données, CLAPI propose une exportation de ses descripteurs dans le format Dublin Core/Olac. Cette

(15) <http://talkbank.org/>

(16) <http://www.univ-orleans.fr/eslo/>

(17) <http://www.projet-pfc.net/>

(18) <http://www.uclouvain.be/valibel-corpus.html>

fonctionnalité a été utilisée notamment pour alimenter le programme « Corpus de la parole »⁽¹⁹⁾ de la DGLF et des Fédérations ILF et TUL. Par ailleurs, l'ensemble des métadonnées de CLAPI et le contenu des transcriptions alignées est disponible en TEI (P5) ; cette solution a été officiellement présentée en 2009 au TEI Council.

En ce qui concerne les interfaces possibles entre banques de données, plusieurs projets sont en cours. Une interface opérationnelle est en cours d'évaluation entre les corpus de la Talkbank en anglais et en danois (corpus *Movin*)⁽²⁰⁾. Dans sa phase de test, CLAPI Workbench intègre 8 corpus de 46 enregistrements (166.000 *tokens*) de la TalkBank, interrogeables avec les outils de requête de la plateforme CLAPI. À terme, l'intérêt de l'interopérabilité est précisément de pouvoir disposer de ces outils pour interroger la totalité des corpus de la TalkBank (qui ne dispose que des moteurs de recherche de CLAN, peu ergonomiques). Dans le cadre du projet CIEL, un test d'intégration dans CLAPI d'un corpus du système MOCA a été effectué en juillet 2008 pour vérifier la compatibilité entre les deux applications. En outre, un projet d'étude « Interopérabilité des grandes banques de données de corpus oraux » est actuellement coordonné par le groupe ICOR pour proposer une solution interopérable entre les banques de données ESLO, PFC et CLAPI.

Face à la prolifération des projets de banques de données, l'interopérabilité est une réponse indispensable, qui nécessite un positionnement clair en matière de formats du signal, de conventions, de logiciels d'alignement, de définition des formats XML, de fonctions d'exportation et d'outils de conversion. Ces derniers permettent de flexibiliser la question de la standardisation, en permettant à la fois des solutions spécifiques aux intérêts analytiques des différentes communautés et des solutions interopérables.

4. L'exploitation analytique des données grâce à des outils d'interrogation dédiés à l'étude de la LPI

Lors du lancement de CLAPI 2.0, un saut considérable a été franchi en passant d'une banque de données archivant des corpus à une plateforme mettant à disposition des outils de recherche : on est ainsi passé d'un ensemble de *corpus archivés* à un ensemble de *corpus outillés*. Ce saut qualitatif est aujourd'hui indispensable dans la perspective de la linguistique de corpus : il est inutile de stocker de grandes masses de données si celles-ci ne sont pas exploitables par des moteurs de recherche. Dans ce sens, plus les masses de données disponibles augmentent et plus les outils d'exploitation doivent être performants.

La base CLAPI est la seule plateforme outillée opératoire dans le domaine du FPI. Deux types de ressources technologiques y ont été implémentés, qui permettent d'une part un balayage automatique des transcriptions et d'autre part des requêtes spécifiques.

(19) Mis en œuvre par la DGLF, l'ILF et le ministère de la culture et de la communication
<http://corpusdelap parole.in2p3.fr>

(20) En ligne à l'adresse <http://clapi.univ-lyon2.fr/Anglais>

4.1. Des outils exploratoires adaptés à la LPI

Le balayage automatique des transcriptions⁽²¹⁾ a été conçu afin d'obtenir des informations sur les caractéristiques principales des données.

Comme il est d'usage dans la linguistique de corpus, orientée principalement vers le texte, ces outils portent sur le lexique : ils permettent d'afficher les attestations d'un item lexical, sa fréquence absolue d'utilisation et ses co-occurrences plus ou moins fréquentes, ainsi que sa localisation dans le tour de parole d'un locuteur, ou au cours de l'interaction considérée dans sa globalité. Selon les besoins de l'analyse, il est toujours possible d'accéder, par *streaming*, au signal audio ou vidéo des attestations ainsi obtenues, et de retourner aux métadonnées des corpus.

Les outils de requête tiennent également compte des phénomènes spécifiquement interactionnels, à savoir, à l'heure actuelle, les pauses, mesurées ou non, et les chevauchements, en distinguant entre segments chevauchés et chevauchants au cours de l'interaction. Il est également possible de combiner ces interrogations avec la recherche d'un item lexical : connaître par exemple la fréquence de la particule discursive *alors* après une pause, ou celle de *attends* dans un segment en chevauchement. Ainsi un outil de recherche de co-occurrences orienté vers les phénomènes interactionnels permet d'identifier les *tokens* les plus souvent présents en début de tour, en début de segment chevauchant/chevauché, après une pause ou dans des tours courts.

En outre, un outil automatique, « Aperçu des traits contextuels d'un *token* », permet de caractériser de manière synthétique la distribution d'un *token* selon les corpus, indiquant s'il est employé le plus souvent dans des interactions en face à face vs dans des polylogues, s'il est employé fréquemment en répétition, en début de chevauchement, autour d'une pause, à une place donnée du tour... dans une sélection de corpus en comparaison avec la base entière.

Enfin, figure au nombre de ces outils une série de caractérisations globales calculées automatiquement pour chacun des enregistrements, concernant le lexique utilisé (hapax, fréquence relative par locuteur), l'enchaînement des prises de parole des interactants, ou la répartition des phénomènes interactionnels au long de l'interaction, et livrée sous forme de graphes.

4.2. Des outils construits par et pour les chercheurs en interaction

Cette seconde série de ressources recourt à des outils permettant des requêtes complexes ciblées. Dans cette section, nous en détaillons trois.

Premièrement, l'outil « Requêtes multi-critères » : il permet de dégager des clusters de phénomènes, en associant des paramètres portant à la fois sur un ou plusieurs *tokens*, pris dans un ordre et à une distance configurables, ou à une certaine position dans le tour de parole, en les combinant avec l'un des phénomènes interactionnels retenus (chevauchements et pauses), et en ajoutant si nécessaire des critères issus des métadonnées (genre interactionnel, nombre de partici-

(21) La recherche portant directement sur le signal sonore n'est encore qu'à l'état de prospection.

pants, etc.). Il est ainsi possible de construire des interrogations « à la carte » : connaître par exemple l'environnement de la particule discursive *ben* située en début de segment chevauchant dans des interactions réunissant au moins quatre locuteurs se déroulant en contexte institutionnel, ou dans celui moins formel de conversations de table entre familiers.

Deuxièmement, l'outil « Détection automatique des segments répétés » : cet outil a évolué grâce à un projet de recherche sur la répétition ⁽²²⁾ ; il permet non seulement de détecter des fragments répétés par un même locuteur ou repris par d'autres, mais d'identifier automatiquement des épisodes caractérisés par de nombreuses répétitions, et de renseigner sur la taille des segments répétés, le lexique concerné, sa fréquence, le locuteur qui utilise le plus cette ressource, d'accepter un ordre différent des items répétés, d'inclure les quasi-répétitions, ou d'exclure certains termes.

Troisièmement, une nouvelle perspective d'exploitation des données se dessine avec les travaux en cours en Recherche d'Informations (RI) ⁽²³⁾, permettant la détection assistée d'épisodes particuliers, comme par exemple des échanges conflictuels, caractérisés par de nombreuses propriétés convergentes : il s'agit de favoriser la recherche de patterns complexes pour assister le linguiste dans l'identification des meso/macro phénomènes dans les transcriptions.

Pour clore cette section, précisons que l'ensemble de ces ressources est à la libre disposition des internautes pour 70% des transcriptions balisées. Quel que soit l'outil utilisé, les résultats des requêtes affichent un contexte significatif de la transcription, un lien vers les descripteurs, et un accès au signal correspondant ; le retour aux données primaires est, en effet, une étape incontournable de la démarche interactionnelle. Ces outils offrent aux chercheurs des aides pour construire des hypothèses de recherche ou pour valider leurs analyses sur un volume significatif de données.

5. Exemples d'utilisations de la plateforme CLAPI

Notre objectif est ici de présenter quelques atouts de la plateforme outillée CLAPI pour la recherche en analyse interactionnelle, et d'esquisser certains usages dans d'autres domaines comme la didactique.

5.1. Applications pour l'analyse interactionnelle

L'approche traditionnelle en analyse des interactions se caractérise par une démarche foncièrement qualitative en ce qu'elle s'attache à la description des ressources langagières mises en œuvre par les participants dans leurs co-productions *in situ*. Si le recours à des « quantifications informelles » (Schegloff 1993)

(22) Il s'agit du projet ANR SPIM, « L'imitation dans la parole : des intégrations sensori-motrices à la dynamique des interactions conversationnelles », dirigé par Noël Nguyen (Laboratoire Parole et Langage, Aix-en-Provence), dont le groupe ICOR est partenaire.

(23) Menés en collaboration avec le laboratoire ERIC, <http://eric.univ-lyon2.fr>, dans le cadre d'un projet Cible financé par la Région Rhône-Alpes.

a toujours été présent en analyse conversationnelle et plus largement dans le champ interactionniste, si la constitution de « collections », ou sous-corpus illustrant un phénomène particulier est une pratique caractéristique de l'approche conversationnelle⁽²⁴⁾, l'augmentation récente de la quantité de données disponibles rend pensable de procéder à de réelles quantifications.

Un des intérêts de CLAPI est de rendre possible une quantification qui est à entendre moins en termes de production de résultats statistiques directement utilisables qu'en termes d'aide à la constitution de collections. La plateforme donne accès à une grande masse de données en vue de la généralisation et de la représentativité des analyses ; elle permet aussi de valider les analyses qualitatives et d'ouvrir des pistes de recherche pour les analyses quantitatives. C'est dans cette articulation entre analyses qualitative et quantitative que nous semble résider le principal intérêt de l'analyse outillée proposée dans CLAPI.

Nous illustrerons cette perspective par l'étude de la particule *voilà*. Le laboratoire ICAR a depuis longtemps développé l'étude des particules discursives en interaction, ou « petits mots » de l'interaction⁽²⁵⁾. Ainsi l'analyse interactionnelle et multimodale de Bruxelles & Traverso (2006) avait distingué trois valeurs pour les usages de *voilà* dans une réunion de conception entre architectes, qui à l'époque n'était pas encore intégrée dans CLAPI :

- a) un emploi de déictique qui, lorsqu'il accompagne une activité de recherche physique ou intellectuelle, traduit un mouvement qualifié de « eurêka » ;
- b) un outil de transition thématique ou actionnelle ;
- c) un marqueur de consensus dans des séquences éminemment collaboratives où *voilà* a souvent une dimension clôturante.

Les auteures concluaient à la nécessité d'étudier « d'autres situations d'interaction (pour) garantir l'extension et la généralisation » de ces descriptions (2006 : 90). Cette piste a donné lieu à deux études ultérieures, basée sur CLAPI.

Deux ans plus tard une analyse menée par le groupe ICOR (2008) sur l'ensemble de CLAPI montre que la répartition des attestations dans les corpus de la base est à première vue très hétérogène puisque, à longueur analogue (environ 60 mn), la fréquence varie de 14 (dans le corpus *Français des années 80* qui réunit des entretiens dans le cadre d'une enquête sociolinguistique) à 90 (corpus *Vente à domicile* d'encyclopédies). La caractérisation du type d'activités déployées dans l'interaction suggère que des critères situationnels pèsent sur la fréquence et les conditions d'utilisation de *voilà* : les manipulations d'objets semblent favoriser son occurrence, alors que la situation d'entretien la réduirait ; le médium utilisé semble également intervenir, puisque on note un nombre plus important de *voilà* au téléphone. Le travail s'est ensuite poursuivi sur un ensemble de corpus présentant une certaine homogénéité avec le corpus *Mosaic*, afin de tester sur une collection plus étendue les trois descriptions de *voilà* proposées. Cette interrogation a confirmé les premières analyses et a surtout permis d'affiner les contextes propres à chacune des valeurs, en mettant en évidence

(24) Cf. sur ce point Schegloff (1996). Voir aussi Mondada (2005).

(25) Cf. en bibliographie, les publications du groupe ICOR pour les dernières publications, et antérieurement Bruxelles & Traverso (2002), Mondada & Traverso (2005), Traverso (2005).

l'importance des conglomérats de particules (56% des attestations de *voilà*), du type de particules et de l'ordre des particules dans ces conglomérats, et enfin de leur place dans le tour : par exemple, *ben*⁽²⁶⁾ en cooccurrence gauche de *voilà*⁽²⁷⁾ relève du type « eurêka » si cette séquence est précédée d'un *ah* ; elle fonctionne comme introducteur de discours indirect au cours d'une narration, donc en milieu de tour ; alors qu'en fin de tour, précédé ou non d'un « et », *ben voilà* est nettement porteur d'une valeur conclusive. D'autres résultats ont aussi pu être validés quant à l'effet de la longueur du tour (les tours courts favorisant l'emploi déictique et affiliatif), et de la localisation de *voilà* (la fin du tour favorisant l'emploi de marqueurs de transition).

La pertinence du placement de *voilà* dans des tours chevauchants nous a conduit à poursuivre l'analyse et à étudier les chevauchements dans CLAPI, et notamment la composition du début des tours chevauchants, en nous focalisant sur la présence de *voilà* dans cette position (ICOR, 2009). Deux grands types d'usage de *voilà* à l'initiale d'un segment chevauchant ont été distingués, caractérisés par le croisement entre un critère fonctionnel – celui de l'action affiliative *vs* disaffiliative assurée par la particule – et un critère formel – un emploi isolé où *voilà* a valeur de tour *vs* un emploi où la particule est suivie d'une expansion.

L'enseignement tiré de cette expérience est double : d'une part, il est évident que l'exploration de grands corpus permet une vérification d'analyses effectuées manuellement sur des corpus plus restreints, ainsi qu'un développement des analyses concernant la distribution à la fois séquentielle et contextuelle des occurrences du phénomène. D'autre part, il apparaît aussi clairement que pour des phénomènes aussi complexes, faisant intervenir des faisceaux de propriétés, les moteurs de recherche constituent une aide indispensable mais ne sauraient remplacer le travail du chercheur dans l'évaluation des résultats et dans l'analyse fine des extraits obtenus en réponse. Cela invite à une démarche mixte, articulant démarche manuelle et semi-automatique, considérant les résultats obtenus comme des pistes heuristiques générant de nouveaux questionnements et pistes d'analyse.

5.2. Autres applications

Bien que conçu au sein d'une démarche interactionniste, CLAPI n'exclut pas des recherches venant d'autres horizons théoriques. En effet, les corpus d'interactions documentées dans des contextes socioculturels différenciés sont sollicités par des disciplines très diverses, allant de l'ethnologie et l'anthropologie à la micro-sociologie et l'*oral history* intéressées par les pratiques communicationnelles, de la psychologie intéressée par la cognition située, distribuée et incarnée, aux sciences de l'éducation intéressées par les processus situés d'apprentissage. En outre, les communautés de praticiens sont aussi très demandeuses de corpus de langue parlée en situation authentique, comme les formateurs, les psychologues et cliniciens du travail, les enseignants et didacticiens axant leur en-

(26) Particule la plus fréquemment associée à *voilà*, en comparaison de *bon*, *alors* ou de morphèmes confirmatifs comme *oui/ouais*.

(27) L'ordre inverse *voilà ben* est nettement plus rare.

seignement sur les pratiques communicatives en contexte et les usages situés de la langue.

Nous esquissons ici quelques pistes inspirées des démarches d'utilisateurs de CLAPI, inscrites dans d'autres disciplines que la linguistique interactionnelle tout en pouvant se combiner avec elle.

Un premier exemple est donné par l'exploitation de l'épaisseur historique de la base. CLAPI documente des interactions depuis 30 ans, ce qui ouvre sur diverses exploitations : la plus immédiate est celle de recherches lexicales comparant des usages d'un mot durant les dernières décennies, éventuellement au sein d'un groupe ciblé de locuteurs (par ex. les jeunes⁽²⁸⁾). D'autres enquêtes, plus originales, concernent l'histoire et le changement des types d'activités – comparant par exemple la manière dont on offre à boire, dont on s'identifie en ouverture d'une conversation téléphonique, dont on présente une consigne à des élèves. Le fait que CLAPI contiennent des conversations à table ou des interactions en classe enregistrées sur trois décennies rend ces comparaisons possibles⁽²⁹⁾.

Un deuxième exemple est celui des exploitations didactiques de CLAPI. Les enseignants de langue seconde voulant introduire leurs élèves à des usages socialement situés de la langue se confrontent souvent à des manuels reconstruisant de manière artificielle des échanges ordinaires – alors que l'utilisation de matériaux authentiques offre bien d'autres sources de réflexion et d'apprentissage.

Une autre exploitation dans le domaine de la didactique comme dans celui de l'acquisition concerne la documentation ponctuelle ou longitudinale d'activités d'apprenants, permettant de développer des études empiriques sur les logiques d'apprentissage/d'acquisition en situation et d'étudier le rôle plus ou moins facilitateur de certains contextes ou certains modes d'interaction⁽³⁰⁾.

6. Conclusions

Pendant longtemps, la linguistique de corpus a essentiellement travaillé sur des corpus de textes écrits, en l'absence de grandes masses de données disponibles pour l'oral. Actuellement, l'essor des corpus d'oral authentique, et encore davantage d'oral en interaction, permet d'envisager une linguistique outillée de l'oral, qui réponde aux principes énoncés par Biber :

- « The essential characteristics of corpus-based analysis are :
- it is empirical, analyzing the actual patterns of use in natural texts ;

(28) Cf. les corpus constitués par ICOR dans le cadre du programme « Corpus de la parole » de la DGLF-LF : « Grammaire, interaction et multimodalité devant la *playstation* » (corpus *jeux vidéo* 2007-08) ; « Comment parle-t-on à 18 ans ? Une enquête sur corpus à propos des pratiques interactionnelles des jeunes de 18 ans » (corpus *conscripts* 2008, et corpus *rassemblement jeunes européens* 2008-09).

(29) Cf. un projet en cours du groupe ICOR sur l'analyse diachronique des interactions financé par le Cluster 13 de la Région Rhône Alpes.

(30) CLAPI héberge des données recueillies par des didacticiens des sciences et des mathématiques. Depuis cette communauté a développé sa propre banque de données de corpus, VISA (consultable sur <http://visa.inrp.fr/visa>).

- it utilizes a large and principled collection of natural texts, known as a « corpus », as the basis for analysis ;
 - it makes extensive use of computers for analysis, using both automatic and [human-computer] interactive techniques ;
 - it depends on both quantitative and qualitative analytical techniques »
- (Biber *et alii*, 1998, 4)

On remarquera la récurrence du terme de « text » utilisé par Biber pour désigner les corpus : les développements actuels permettent précisément de dépasser une linguistique de corpus qui se fonde sur le texte pour développer ses modèles et ses instruments – en l'absence d'outils qui rendent la complexité des corpus oraux appréhendable de manière informatisée.

Cet article a voulu présenter certaines de ces solutions, qui permettent à la fois de relever un tel défi et de respecter un certain nombre de principes souvent jugés comme incompatibles : d'une part la spécificité de l'oral en interaction, telle qu'elle est observable dans la complexité des données primaires et des données secondaires, ainsi que le caractère situé, indexical, incarné de la parole en interaction ; d'autre part les exigences de standardisation, d'homogénéisation des formats et des descriptions. L'enjeu et le défi consistent précisément à sauvegarder la complexité de l'objet tout en imaginant des solutions techniques qui en permettent une exploitation outillée. La démarche qui anime la plateforme CLAPI repose sur un tel pari.

Les défis à venir concernent précisément des dimensions constitutives de la complexité des corpus de parole en interaction. Le premier défi porte sur le traitement semi-automatique de la multimodalité, qui présuppose une standardisation des modes d'annotation, non encore atteinte, alors même que les différents outils d'alignement et d'annotation permettent aujourd'hui des conversions entre un système et l'autre (entre CLAN, ELAN, ANVIL, Exmaralda, Praat...). Le second défi est posé par le traitement de corpus plurilingues ou polylectaux : le projet CIEL vise à développer des analyses de la variation de la grammaire du français dans différentes aires de la francophonie et dans différents types d'activités interactionnelles, et suppose donc que l'on puisse traiter de variétés du français qui sont parfois très éloignées les unes des autres. La maîtrise de cette hétérogénéité sans la réduire à des catégories trop simples, ainsi que la reconnaissance des catégories des locuteurs tout en implémentant des balises standardisées nécessaires pour les requêtes, définissent un équilibre difficile à tenir.

Ces nouveaux défis montrent l'imbrication fondamentale des questions théoriques et analytiques avec les questions techniques et technologiques, qui est un trait définitoire des entreprises de corpus. Cela motive une démarche linguistique qui intègre le plus possible les différentes dimensions constitutives des corpus : de la méthodologie de terrain aux dispositifs technologiques d'enregistrement, des formats de stockage aux questions d'anonymisation des données, des standards des conventions de transcription aux représentations de la spécificité de la parole située et incarnée.

Bibliographie

- BALTHASAR, L. & BERT, M. (2005) : « La plateforme Corpus de Langues Parlées en Interaction (CLAPI) : Historique, état des lieux, perspectives », *Lidil* 31, 13-33.
- BANGE, P. (éd.) (1987) : *L'analyse des interactions verbales. La Dame de Caillure : une consultation*, Berne, Peter-Lang.
- BAUDE, O. (éd.) (2006) : *Corpus oraux. Guide des bonnes pratiques*. Paris : CNRS éditions.
- BIBER, D., CONRAD, S., & REPPEN, R. (1998) : *Corpus Linguistics : Investigating Language Structure and Use*. Cambridge : Cambridge University Press.
- BRUXELLES, S. & TRAVERSO, V. (2002) : « Les corpus de langue parlée en interaction au GRIC » ; in : Pusch, K. & Raible, W. (eds) : *Romanistische Korpuslinguistik*. Tübingen : Narr Verlag, 59-70.
- (2002) : « Ben : apport de la description d'un "petit mot" du discours à l'étude des polylogues », *Marges linguistiques* 2.
- (2006) : « Usages de la particule *voilà* dans une réunion de travail : analyse multimodale » ; in : Drescher, M. & Job, B. (eds) : *Les marqueurs discursifs dans les langues romanes : approches théoriques et méthodologiques*. Bern : Lang, 71-92.
- BRUXELLES, S., MONDADA, L., SIMON, A.-C., TRAVERSO, V. (2009) : « Grands corpus de français parlé : bilan historique et perspectives de recherches », *Cahiers de Linguistique de Louvain* 33/2, 1-14.
- COSNIER, J. & KERBRAT-ORECCHIONI, C. (éds) (1986) : *Décrire la conversation*. Lyon : PUL.
- DÉTIENNE, F. & TRAVERSO, V. (éds) (2009) : *Méthodologies d'analyse de situations coopératives de conception : Corpus Mosaic*. Nancy : PUN.
- DISTER, A., GADET, F., LUDWIG, R., LYCHE, C., MONDADA, L., PFÄNDER, S., SIMON, A.C., SKATTUM, I. (2008) : « Deux nouveaux corpus internationaux du français : CIEL-F et CFA ». *Revue de Linguistique Romane*, vol. 72, 295-314.
- ETIENNE, C. (2009) : *La TEI dans le projet CLAPI. Corpus de Langues Parlées en Interaction*. TEI Council, Lyon - avril 2009.
- GADET, F., LUDWIG, R., PFÄNDER, S. (2008) : « Francophonie et typologie des situations », *Cahiers de Linguistique*, vol. 34/2, 143-162.
- GROUPE ICOR (L. BALTHASAR, S. BRUXELLES, L. MONDADA et V. TRAVERSO) (2007) : « Variations interactionnelles et changement catégoriel : l'exemple de "attends" » ; in *Langues, Cultures, Interaction. Actes du Colloque du Réseau Français de Sociolinguistique*, 2005. Paris : L'Harmattan, 299-320.
- GROUPE ICOR (BERT, M., BRUXELLES, S., ETIENNE, C., MONDADA, L., TESTON, S., TRAVERSO, V.) (2008) : « "Oh::, oh là là, oh ben..." , les usages du marqueur "oh" en français parlé en interaction » ; in Durand, J., Habert, B., Laks, B. (éds.), *CMLF (Congrès mondial de linguistique française)*. Paris : France (9-12 juillet 2008) - <http://www.linguistiquefrancaise.org/index.php?option=toc&url=/articles/cmlf/abs/2008/01/contents/contents.html>
- GROUPE ICOR (M. BERT, S. BRUXELLES, C. ETIENNE, L. MONDADA, V. TRA-

- VERSO) (2008) : « Tool-assisted analysis of interactional corpora : “voilà” in the CLAPI database », *Journal of French Language Studies*, 18/01, 121-145.
- GROUPE ICOR (M. BERT, S. BRUXELLES, C. ETIENNE, L. MONDADA, V. TRAVERSO) (2009) : « Exploitation de la plateforme Corpus de Langue Parlée en Interaction (CLAPI) : le cas de “voilà” dans les chevauchements », *Cahiers de Linguistique de Louvain* 33/2, 243-268.
- JEFFERSON, G (2004) : « Glossary of transcript symbols with an introduction » ; in Lerner, G. H. (ed.) : *Conversation Analysis : Studies from the first generation*. Amsterdam : Benjamins, 13-31.
- MONDADA, L. (2005) : « L’analyse de corpus en linguistique interactionnelle : de l’étude de cas singuliers à l’étude de collections », in. A. Condamine (éd.), *Sémantique et corpus*, Paris : Hermès, 76-108.
- (2006) : « Video Recording as the reflexive Preservation of Fundamental Features for Analysis » ; in Knoblauch, H., Raab, J., Soeffner, H.-G., Schnetler, B. (eds). *Video Analysis*, Bern : Lang, 51-68.
- (2007) : « Enjeux des corpus d’oral en interaction : re-temporaliser et resituer le langage », *Langage et Société*, 121-122, 143-160.
- (2008) : « Documenter l’articulation des ressources multimodales dans le temps : la transcription d’enregistrements vidéos d’interactions » ; in M. Bilger (éd). *Données orales, les enjeux de la transcription*. Perpignan : Presses Universitaires de Perpignan, 127-155.
- MONDADA, L. & TRAVERSO, V. (2005) : « (Dés)alignements en clôture » *Lidil* 31, 35-59.
- SACKS, H. (1984) : « Notes on methodology » ; in Atkinson, J.M. & Heritage, J. (ed.), *Structures of social action : Studies in Conversation Analysis*. Cambridge : Cambridge University Press, 21-27.
- SCHEGLOFF, E. A. (1993) : « Reflections on Quantification in the Study of conversation », *Research on Language and Social Interaction*, 26/1, 99-128.
- (1996) : « Turn organization : One intersection of grammar and interaction », in Ochs, E., Schegloff, E.A. & Thompson, S.A. (eds.), *Grammar and Interaction*. Cambridge : Cambridge University Press, 52-133.
- SELTING, M. *et al.* (1998) : « Gesprächsanalytische Transkriptionssysteme (GAT) ». *Linguistische Berichte*, 173, 91-122.
- TRAVERSO, V. (2005) : « Quelques formats intégrant la répétition comme ressource pour le développement thématique dans la conversation ordinaire », *Rivista di Psicolinguistica Applicata*, IV-2/3, 153-166.