Lidil
Revue de linguistique et de didactique des langues

31 | 2005
Corpus oraux et diversité des approches

La plateforme « Corpus de langues parlées en interaction » (CLAPI)

Historique, état des lieux, perspectives.

Lukas Balthasar et Michel Bert

revues.org

Édition électronique
URL : http://lidil.revues.org/139
ISSN : 1960-6052

Éditeur
Ellug / Éditions littéraires et linguistiques de l'université de Grenoble

Édition imprimée
Date de publication : 1 juin 2005
Pagination : 13-33
ISBN : 2-914176-12-0
ISSN : 1146-6480

Référence électronique
Lukas Balthasar et Michel Bert, « La plateforme « Corpus de langues parlées en interaction » (CLAPI) », *Lidil* [En ligne], 31 | 2005, mis en ligne le 03 octobre 2007, consulté le 30 septembre 2016.
URL : http://lidil.revues.org/139

La plateforme « Corpus de langues parlées en interaction » (CLAPI)

Historique, état des lieux, perspectives.

Lukas Balthasar et Michel Bert

1 Les sciences du langage s'appuient de plus en plus sur des traitements informatisés de corpus écrits et, depuis peu, oraux. Dans le contexte de l'oral tout particulièrement, la mise à disposition d'un corpus suppose un traitement préalable des données permettant par la suite divers types d'interrogations et d'usages informatiques. Alors que l'on dispose aujourd'hui de bases de données orales importantes aux États-Unis, mais aussi en Angleterre, en Allemagne ou en Espagne, ces ressources sont encore très insuffisamment développées en France[1]. Ce type de données est pourtant fondamental tant pour le développement de la linguistique, des sciences sociales et de l'histoire que d'un point de vue patrimonial.

2 La plateforme *Corpus de langues parlées en interaction* (CLAPI)[2] du laboratoire ICAR est un environnement d'archivage et d'analyse de corpus d'interactions enregistrées en situation authentique (en famille et entre amis, sur le lieu de travail, dans les institutions les plus diverses, etc.). Dans son état actuel, la plateforme CLAPI comporte environ 75 corpus, soit plus de 150 heures de données transcrites ou 2,5 millions de mots, et fournit un ensemble croissant d'outils pour la construction de ce type de corpus, pour leur stockage, leur communication et, surtout, pour leur analyse qualitative et quantitative[3]. L'orientation scientifique de CLAPI est d'abord déterminée par l'une des spécialités scientifiques traditionnelles du laboratoire ICAR, l'analyse interactionniste de faits de langue et de discours en contexte[4]. En outre, dans la mesure où CLAPI rassemble les fruits de deux décennies de recherches menées par plusieurs laboratoires successifs d'orientation pluridisciplinaire – mais aussi grâce aux compétences des divers partenaires nationaux et internationaux –, le contenu et la conception de la base pourront également répondre aux besoins issus d'autres disciplines et courants théoriques en sciences humaines[5]. Le nombre, la variété et la qualité des corpus rassemblés dans la base CLAPI lui confèrent aujourd'hui une valeur scientifique spécifique qui la distingue non

seulement d'autres types de corpus, écrits et oraux, mais qui paraît aussi particulièrement prometteuse pour le développement de nouvelles approches et méthodes en linguistique interactionnelle et en pragmatique au sens large du terme. Le caractère authentique et la profondeur historique des corpus CLAPI confèrent également à la base une valeur patrimoniale unique dans une culture où domine la valorisation du langage écrit et de l'œuvre littéraire. Le fonctionnement de la base clapi a initialement été assuré par les projets du laboratoire en collaboration avec un certain nombre de partenaires. Selon les moyens disponibles, d'autres potentialités orientées vers la communauté scientifique plus large seront réalisés dans les années à venir.

3 Le développement d'un projet comme la plateforme CLAPI dépend de la conjonction de facteurs scientifiques, technologiques et institutionnels. Ainsi, le développement des théories interactionnistes à partir des années 60 a été rendu possible par l'émergence du magnétophone portable. L'usage des supports audio-visuels, relativement complexes et coûteux (et donc souvent réservé à des structures de recherche d'une certaine taille), a encore longtemps été cantonné à des situations expérimentales. Les membres des structures dont est issue l'UMR ICAR actuelle, notamment le GRIC, ont fait un travail pionnier dans ce domaine en France, non seulement en introduisant les approches interactionnistes dès les années 75, mais aussi en créant un contexte institutionnel dans lequel un grand nombre de corpus audio et audio-visuels naturels et expérimentaux a pu être réalisé[6].

4 Ces corpus jusque-là dispersés constituent aujourd'hui le fondement de la base CLAPI. Depuis la fin des années 90, les corpus CLAPI font l'objet d'un vaste programme de numérisation afin de garantir leur pérennité, de permettre un traitement unifié et de faciliter leur circulation entre chercheurs. Cela a donné lieu à la constitution d'une Médiathèque de corpus, composée de plus de 250 CD-ROMs. La vaste majorité des transcriptions existantes (doc, word, txt, etc.) sont converties vers un format XML spécialement adapté à la structure et aux fonctions des moteurs de recherche intégrés dans la plateforme CLAPI. Les principaux phénomènes interactionnels transcrits – productions verbales, chevauchements, pauses, etc. – sont balisés et reliés avec le signal audio ou vidéo correspondant.

5 Dans le cadre d'une collaboration interdisciplinaire[7], la base de données CLAPI a été conçue dans la perspective de répondre à un certain nombre d'exigences – complexité, robustesse, accessibilité, extraction – qui permettront la gestion d'un nombre important de corpus, leur description précise au moyen d'un ensemble de descripteurs (*attributs* en terme informatique), des requêtes sur corpus (descripteurs et transcriptions alignées avec les enregistrements), la mise en place de procédures d'intégration en ligne de corpus constitués à l'intérieur ou à l'extérieur de l'*UMR ICAR*, la gestion de la consultation des corpus par des droits d'accès. Durant l'année 2004, un prototype développé en collaboration avec le laboratoire ERIC a été implémenté en PHP/MySQL pour tester la sécurisation de l'accès aux données référencées dans CLAPI.

6 Dans le même temps, les campagnes de construction de corpus se poursuivent, d'une part pour enrichir certains fonds existants qui sont particulièrement importants pour la recherche en linguistique interactionnelle – ou pour leur valeur patrimoniale –, d'autre part pour intégrer de nouveaux types de corpus permettant de traiter certaines questions liées au développement de la base (par exemple aspects juridiques et éthiques, analyse de données vidéo, etc.)[8].

7 Les caractéristiques de l'oral, les orientations interactionniste et sociolinguistique, ainsi que la méthodologie qualitative impliquent que la notion de *corpus* se distingue en partie de celle qui a cours en linguistique de corpus[9]. Dans le cadre de la plateforme CLAPI, un corpus est un ensemble composé a) d'enregistrements audio ou vidéo d'interactions et b) de leurs transcriptions, éventuellement accompagnés c) de documents annexes. La constitution des corpus vise en priorité à rendre accessibles la qualité et la complexité des phénomènes de la langue parlée dans des contextes authentiques ; les considérations de représentativité quantitative et de traitement informatique des données n'intervenant, par contre, que dans un second temps[10]. L'unité d'un corpus est définie par une certaine *homogénéité* qui peut provenir des caractéristiques de la situation d'enregistrement (sites, terrains, etc.), des caractéristiques interactionnelles et linguistiques des données enregistrées (genre, type d'activité, etc.) ou des caractéristiques des interactants (leur compétence d'apprenant, par exemple).

8 CLAPI est constitué d'une très grande variété de corpus individuels ou collectifs, notamment un fonds croissant de *conversations ordinaires* entre membres d'une famille, entre amis, voisins, etc., considérées comme prototypique de l'usage du langage en linguistique interactionnelle (voir par exemple Sacks, Schegloff & Jefferson, 1974 ; Schegloff, 1992 ; Traverso, 1996 ; Mondada, 2001). En France, un important corpus audio de ce type, *Conversation familière*, a été constitué à Lyon à la fin des années 80 (Traverso, 1996, 1999, à paraître). CLAPI comporte probablement l'un des premiers corpus vidéo expressément construit pour des analyses linguistiques/interactionnistes : le corpus *Mode* , composé de trois conversations avec thème, durée et interlocuteur imposés, a lui aussi donné lieu à une série de publications importantes (voir par exemple Cosnier & Kerbrat-Orecchioni, 1987). Dans le cadre des réflexions menées au sein de CLAPI et des projets de recherche actuels, une campagne a été lancée afin de constituer un corpus vidéo de *Conversations ordinaires* qui réponde à des exigences élevées en termes de construction du terrain, de dispositifs d'enregistrement audio-visuel, de préparation des données, d'archivage et d'exploitation scientifique (Balthasar & Mondada, 2003-...).

9 Les corpus hébergés dans CLAPI ont été recueillis dans des contextes de conversations ordinaires, mais également dans différents types d'institutions, de services publics ou d'entreprises privées (poste, mairies, études notariales, commerces, etc.), dans l'enseignement secondaire et universitaire, dans différents contextes médicaux (cabinet, hôpital), dans les médias et sur le Web (*les chats*). Sont également hébergés dans CLAPI des corpus réalisés dans le cadre d'enquêtes dialectologiques (fonds en occitan et en francoprovençal de l'Institut Pierre Gardette) et en sociolinguistique, le corpus FRA 80 par exemple[11]. Aujourd'hui, la majorité des corpus est encore enregistrée en audio avec, comme exception majeure, un ensemble important de corpus vidéo réalisés à partir des années 90 par le groupe de recherche COAST. Il est composé principalement d'enregistrements recueillis dans des classes de sciences au lycée[12]. De plus en plus de corpus ne provenant pas d'ICAR sont également hébergés dans CLAPI, par exemple un certain nombre de corpus de français parlé du laboratoire LIDILEM, ainsi que des corpus d'apprenants L1 et L2 développés dans le laboratoire DDL.

10 Dans la base CLAPI, les corpus sont archivés sous forme d'unités documentaires qui comportent obligatoirement :

- une fiche descriptive composée des descripteurs CLAPI ;
- les enregistrements (données primaires) ;

- leurs représentations, en particulier sous forme de transcriptions, accompagnées des conventions correspondantes (données secondaires) ;
- la bibliographie des études portant sur le corpus.

11 À ces éléments peuvent s'ajouter :

- des échantillons de données primaires et secondaires ;
- des documents annexes numérisés (ex. artefacts produits par les interactants, documents utilisés lors de l'interaction, productions du collecteur / notes de terrain, photos, etc.) ;
- des publications portant sur le corpus.

12 Lors de l'identification et du chargement du corpus, le responsable définit lui-même la structure de son corpus, explicite les critères retenus et précise ses liens éventuels avec d'autres corpus. La structure d'une unité documentaire dans CLAPI est hiérarchisée à partir de la fiche descriptive : un corpus comporte un ou plusieurs enregistrements, accompagnés éventuellement de documents annexes. Chaque enregistrement est associé à une ou plusieurs transcriptions (ex. versions successives anonymisées ou non, de formats différents – Word, Praat, CLAN, etc.) reliées systématiquement à leurs conventions de transcription. Des images annexes peuvent également être stockées dans CLAPI avec les transcriptions (ex. scans de transcriptions manuscrites, etc.). Si le responsable le souhaite, les interactants sont identifiés et leurs principales caractéristiques sociolinguistiques sont précisées dans les descripteurs.

13 Le traitement des corpus réunis dans la plateforme CLAPI présuppose un certain nombre de formats standard, notamment en ce qui concerne la caractérisation des corpus. L'identification de corpus collectés en fonction de finalités scientifiques multiples et diverses est relativement complexe, d'autant plus que les standards informatiques développés jusqu'à présent – par exemple la *Dublin Core Metadata Initiative* – ne sont pas encore suffisamment adaptés aux besoins du traitement des corpus oraux, notamment en linguistique interactionnelle. En adaptant donc l'identification des corpus de la manière la plus précise possible aux exigences de la plateforme CLAPI et en attendant les résultats des initiatives en cours comme ISLE, développées en collaboration avec des linguistes, la liste actuelle des descripteurs CLAPI comporte 75 entrées hiérarchisées (génériques ou spécifiques, obligatoires ou facultatives) couvrant les points suivants[13] :

Informations générales	nom du corpus, texte de présentation, dates et lieux de recueil, liste des corpus associés, durée totale...
Auteurs	responsable, collecteurs, transcripteurs...
Genre interactionnel	interaction privée, de travail, médicale...
Enregistrements	nom, date et lieu de recueil, durée, type de support (audio vidéo), anonymisation...
Transcriptions	convention et logiciels utilisés, orthographe standard ou adaptée, alignement et balisage, exhaustivité, anonymisation...
Locuteurs	nom ou pseudonyme et caractérisation sociolinguistique

Bibliographie	références des travaux réalisés sur le corpus

Conditions de diffusion des enregistrements et des transcriptions : sans limitation, dans le cadre de la signature d'une convention, non accessible.

14 La structure informatique de la base permettra l'évolution de CLAPI en fonction de l'émergence de nouveaux besoins ou de l'adaptation à des *metadata standards* tels que ISLE.

15 La situation concernant les conventions de transcription et les formats d'archivage est comparable à celle rencontrée pour les descripteurs de corpus. L'objectif d'un traitement unifié de la grande variété des corpus intégrés dans la plateforme CLAPI présuppose des formats d'archivage standardisés. Pour des raisons diverses, les conventions de transcription n'ont pourtant pas été unifiées jusqu'à présent en linguistique interactionnelle (comme c'est le cas, par exemple, pour l'API). Dans la mesure où les standards existant pour la représentation informatique de l'oral (TEI, CES, etc.)[14] ne sont pas encore complètement stabilisés et comme ils ne répondent pas actuellement, de manière satisfaisante, aux besoins d'analyse et de représentation des données en linguistique interactionnelle, il s'est avéré inévitable de recourir à des développements réalisés au sein du laboratoire.

16 Ces développements ont abouti à une première version d'une convention de transcription propre au projet CLAPI, la Convention ICOR, ainsi qu'à des prototypes de formats informatiques en XML. L'objectif de la convention a consisté à répondre de la manière la plus précise possible aux exigences de la transcription en linguistique interactionnelle. La convention ICOR tient donc compte d'un certain nombre de principes de transcription généraux, ainsi que de l'état d'avancement du domaine en ce qui concerne le choix, la définition et la représentation (signes de notation) des catégories analytiques retenues. Tout en respectant la représentation traditionnelle des données en linguistique interactionnelle, la convention favorise l'usage des outils de transcription les plus répandus dans le domaine, notamment Praat et CLAN[15]. Les catégories de notation actuellement prises en compte au niveau de la totalité des corpus CLAPI sont les phénomènes de base du discours oral :

- production verbale/tour de parole ;
- chevauchement ;
- pause/silence ;
- token/mot ;

17 ainsi que l'interactant auquel ces phénomènes sont attribués.

18 La convention ICOR comporte en plus une vingtaine de conventions concernant des phénomènes segmentaux, supra-segmentaux et non verbaux (gestuels notamment). Aujourd'hui, la convention ICOR est notamment utilisée pour la réalisation des nouveaux corpus dans les logiciels Praat et CLAN, ainsi que pour l'affichage des résultats produits au moyen des moteurs de requête implémentés dans la plateforme CLAPI[16].

19 Le format XML actuellement développé pour CLAPI vise d'une part une interprétation fidèle de la convention ICOR et, d'autre part, la traduction la plus précise possible des cinq phénomènes cités ci-dessus tels qu'ils sont transcrits dans les corpus de la plateforme[17].

20 La convention ICOR, ainsi que le format XML de la plateforme évolueront nécessairement dans les années à venir en fonction des développements de la linguistique

interactionnelle et des évolutions technologiques, ainsi que des objectifs analytiques des futurs projets de recherche du laboratoire.

21 Pour être intégré dans la base de données, un corpus subit un traitement décomposable en plusieurs étapes successives.

22 – Recueil des divers éléments du corpus et des autorisations signées par les personnes enregistrées.

23 – Numérisation des données primaires et secondaires selon des formats prédéfinis. Cette standardisation s'appuie sur des évaluations portant sur l'accessibilité, le niveau de compression et la pérennité des formats vidéo, audio, texte et image.

24 – Préparation du corpus :

- évaluation de l'application de la convention de transcription à partir d'échantillons ;
- éventuellement, révision de la transcription. Il s'agit en général d'un « toilettage » léger soumis à l'approbation du responsable du corpus ;
- pour les corpus non alignés (txt, rtf, doc...), minutage de la transcription ;
- alignement d'extraits plus ou moins conséquents (au minimum 1 minute).
- éventuellement, conversion des transcriptions vers des formats exploitables par des moteurs de requête.

25 – Anonymisation. Le stockage et la diffusion de certains corpus requièrent l'anonymisation des transcriptions, le bippage des données audio ou même des transformations de l'image vidéo. Les critères d'anonymisation sont établis par le responsable du corpus en tenant compte des recommandations issues de réflexions menées à ICAR en collaboration avec des juristes du CECOJI[18].

26 – Mise en forme de transcriptions alignées avec l'extrait audio ou vidéo correspondant. Ces échantillons seront accessibles en ligne sur CLAPI sans restriction.

27 Gravure d'un ou plusieurs CD-ROMs. Un double du corpus numérisé est rendu au responsable. Les données originales ou intermédiaires (p. ex. transcription ou données audio non anonymisées) sont également parfois conservées dans la Médiathèque de corpus. Celle-ci contient en outre certains documents papier (originaux des autorisations signées par les participants, documents collectés lors de l'enregistrement...).

28 – Identification et intégration du corpus dans la base de données CLAPI.

29 L'intégration de corpus nombreux et très divers a permis d'évaluer le temps de traitement des différentes étapes. Selon la nature du corpus ou le format des données primaires, par exemple, la durée de traitement peut varier du simple au décuple. À l'avenir, il est envisagé que le dépositaire d'un corpus prenne en charge une partie ou la totalité du traitement, avec l'assistance de l'équipe de la Médiathèque. A cette fin, des documents-supports à l'intégration seront mis à disposition : convention/accord pour l'enregistrement et la diffusion, recommandations techniques pour la numérisation, guide d'anonymisation, manuels pour les logiciels de transcription ou d'édition de corpus, aide à l'identification et à l'hébergement d'un corpus dans la base CLAPI.

30 Afin d'assurer la gestion et la sécurisation des accès en ligne aux données hébergées dans CLAPI, le groupe ICOR a identifié les différents acteurs susceptibles d'intervenir dans la base et leurs droits respectifs sur les corpus. Cette réflexion a permis d'établir six profils différents en fonction des possibilités qui leur sont offertes :

31 – *Anonyme*. Il s'agit des personnes ne possédant pas de droit d'accès particulier. Elles peuvent consulter la fiche descriptive des corpus non verrouillés et accéder à leurs

échantillons. Elles peuvent également consulter les corpus en accès libre et effectuer des requêtes sur les données non verrouillées.

32 – *Responsable*. La personne ayant déposé un corpus dans CLAPI est responsable de ses données. Elle signe une charte qui précise les conditions d'hébergement des corpus. Si son corpus n'est pas en accès libre, le responsable peut attribuer des droits d'accès à des tiers, en concertation avec le conseil de gestion de CLAPI. La consultation d'un corpus en accès limité exige la signature d'une convention de recherche entre le responsable du corpus, l'emprunteur et le directeur de l'*UMR ICAR*. L'accès au corpus suppose un échange (réalisation de transcriptions, d'annotation, etc.). L'emprunteur devra également signaler toute publication portant sur le corpus.

33 Un responsable peut verrouiller son corpus – il est alors invisible aux anonymes et aux contractuels – et il peut demander sa suppression. Il est le seul à pouvoir déverrouiller un corpus qu'il aurait lui-même verrouillé. S'il valide les modifications ou ajouts qui lui sont proposés, il devient alors responsable de ces nouveaux éléments.

34 Le responsable peut également effectuer des requêtes sur son propre corpus, si celui-ci a subi le traitement nécessaire. Il précise alors les parties du corpus qui sont accessibles aux requêtes par les anonymes. Il a également été prévu de pouvoir déposer un corpus dans CLAPI tout en le verrouillant, pour pouvoir profiter des outils de requête de la plateforme sans que le corpus soit visible par les anonymes.

35 – *Contractuel*[19]. Une personne ayant signé une convention de recherche avec un responsable devient contractuel. Il peut consulter ou télécharger le corpus ou certains de ses éléments, il peut effectuer des requêtes sur ces données.

36 – *Equipe Médiathèque*. L'équipe Médiathèque est chargée du traitement des corpus en vue de leur intégration dans la base de données et de leur archivage dans la Médiathèque. Elle assure également une assistance technique auprès des dépositaires de corpus. Les membres de l'équipe Médiathèque vérifient que les données figurant dans la base sont conformes aux principes techniques qui la régissent. Ils ont donc accès à l'intégralité des données et ils peuvent verrouiller un élément (corpus, enregistrement, transcription...). Les données verrouillées (nouveau corpus, ajout ou modification validés par le responsable, éléments verrouillés par l'équipe elle-même ou par le conseil scientifique) doivent subir la validation technique de l'équipe Médiathèque avant d'être déverrouillées par le conseil de validation.

37 – *Conseil de validation*. Ses membres évaluent les demandes de dépôt de nouveaux corpus et ils valident les modifications ou ajouts proposés par le responsable. Ils peuvent verrouiller un corpus ou certains de ses éléments. Tous les éléments verrouillés doivent en dernier lieu être validés par le conseil pour être accessibles aux anonymes et aux contractuels.

38 – *Conseil de gestion*. Il est composé de membres du groupe ICOR. Le conseil de gestion gère l'attribution des droits d'accès et la signature des conventions de recherche. Il peut suggérer les modalités d'échange entre responsable et emprunteur. Le conseil de gestion peut éventuellement s'opposer à une demande d'emprunt (par exemple demandes d'accès portant sur une partie trop importante des données de la base CLAPI). Le conseil de gestion peut consulter les données non verrouillées et émettre un avis sur les aspects juridiques concernant un corpus.

39 L'organisation des droits d'accès dans la base s'explique par l'orientation générale donnée au projet : la plateforme CLAPI a été conçue comme un lieu d'hébergement de corpus, et

comme un outil d'analyse à destination des responsables de corpus et des chercheurs à qui les responsables ont attribué des droits à leurs données.

40 En attendant que la base de données puisse assurer la circulation en ligne contrôlée des corpus, le prêt reste expérimental ; il est nominatif et gratuit. Après l'examen de la demande de prêt et la signature de la convention, des CD-ROMs sont envoyés aux emprunteurs.

41 Dans leur vaste majorité, les corpus rassemblés dans la base CLAPI ont été construits dans la perspective d'études qualitatives portant a) sur des *cas particuliers* de structures interactionnelles, tels que, par exemple, le prolongement d'un tour avec des moyens syntaxiques ou gestuels afin de conserver le droit à la parole, b) sur des *collections de structures* permettant de dégager, par exemple, le fonctionnement de l'organisation préférentielle déterminant les phénomènes de réparation ou encore c) sur le fonctionnement de *phénomènes socio-cognitifs* comme le « face-work » dans le cas d'une confidence. L'un des objectifs de la conception de la base CLAPI consiste à favoriser ce type d'analyse[20] au moyen d'outils informatiques aujourd'hui disponibles ou à développer en interne.

42 Afin d'atteindre cet objectif, les nouveaux corpus CLAPI ont été construits en utilisant des outils d'alignement des transcriptions avec le signal (CLAN et Praat notamment). Mis à part leur intérêt pour la réalisation de la transcription, ils fournissent des fonctions analytiques spécifiques, ainsi que d'excellentes interfaces de présentation des données. En permettant de travailler en même temps sur le texte de la transcription *et* sur le signal, le caractère sélectif de toute transcription et la richesse particulière des données audiovisuelles restent continuellement transparents. De plus, des caractéristiques autrefois difficilement accessibles tels que les traits prosodiques des productions verbales ou encore le détail de structures audio-kinésiques peuvent être analysés efficacement avec ces outils.

43 Les outils de requête intégrés dans la plateforme CLAPI – CLAPI QUERY et NXT Search notamment –[21] faciliteront considérablement les analyses qualitatives en donnant un accès très rapide aux phénomènes de base des interactions (les productions verbales/ tours de parole, les chevauchements, les pauses, les token/mots et la temporalité de leur production). Dans la mesure où ces phénomènes sont fondamentaux pour l'organisation des interactions et dans la mesure où les catégories retenues permettent de spécifier avec précision une grande variété de structures assez complexes, les outils de recherche seront d'une grande utilité, d'autant plus qu'ils permettent d'inclure des informations sociolinguistiques dans les requêtes elles-mêmes (et non seulement au niveau préalable de la sélection des corpus sur lesquels porteront ces requêtes). Les outils permettent à l'utilisateur d'effectuer des recherches sur une sélection de corpus CLAPI d'une part et sur des constellations complexes de phénomènes transcrits d'autre part, à savoir par exemple :

- dans des corpus CLAPI de dilogue au téléphone, successions d'au moins quatre productions verbales/tours (PV) réalisées par des interactants différents et séparées par des pauses longues ;
- dans des corpus CLAPI de dilogue au téléphone, les PV réalisées en chevauchement avec au moins cinq token/mots dans chaque PV à partir du chevauchement ;
- dans un sous-ensemble d'un corpus Z, toute production verbale du participant A contenant les tokens « ici », « là », « dessus », « dessous » ou « côté » suivies d'une production verbale

du participant B qui commence par le token « non » et qui se trouve à une distance maximale de 2 PV de la PV de A.

44 La présentation des résultats sera affichée sous forme de listes de segments transcrits et alignés avec le signal. Elle pourra être visualisée dans la convention ICOR ou suivant la convention d'origine.

45 L'usage des moteurs de recherche actuellement développés permet d'élargir le cadre méthodologique de la linguistique interactionnelle et d'y intégrer des méthodes quantitatives. Ces moteurs de recherche permettront ainsi d'effectuer des comptes et des calculs de fréquence tels que :

- la fréquence des PV successives sans pause ni chevauchement par rapport aux PV réalisées en chevauchement et par rapport aux PV séparées par des pauses ;
- la fréquence relative de productions verbales du locuteur A chevauchant une production du locuteur B par rapport au cas inverse (productions verbales du locuteur B chevauchant une production du locuteur A), ainsi que le nombre moyen de tokens contenus dans les productions verbales de A et de B après le chevauchement ;
- une concordance KWIC[22] du token « bof » portant sur l'ensemble de la base CLAPI avec comme contexte cinq tokens à gauche et à droite de « bof » ;
- la fréquence relative du token « oui » par rapport au token « ouais » dans les productions verbales d'interactants âgés de moins de 10 ans, 20 ans, 30 ans... dans l'ensemble des corpus francophones de CLAPI.

46 Bien que le nombre de phénomènes pris en compte soit aujourd'hui encore relativement réduit, il est évident qu'un certain nombre de questions de quantification qui, en linguistique interactionnelle, ont jusqu'à présent été traitées en termes de *quantification informelle* et de *récurrence*, pourront recevoir une interprétation et une validation en terme de *quantification formelle* (à ce sujet, voir par exemple Schegloff, 1993, Balthasar & Mondada, 2003). Même si les réserves vis-à-vis des méthodes quantitatives formulées dans les approches interactionnistes se justifient dans de nombreux cas, certaines hypothèses concernant notamment le système d'organisation des tours de parole ou des organisations préférentielles globales comme celles qui déterminent les réparations pourront graduellement être vérifiées.

47 La plateforme CLAPI a été conçue pour pouvoir héberger un grand nombre de corpus de formats très divers, anciens ou récents, et pour permettre leur exploitation outillée. La structure de la base de données connaîtra encore des évolutions, par exemple au niveau de l'enrichissement et de l'adaptation des descripteurs, des formats de représentation de données finement annotées ou du nombre des phénomènes pris en compte par les outils d'analyse. La base évoluera aussi en fonction des besoins issus de nouveaux horizons disciplinaires. Les expériences de gestion d'une plateforme de corpus de langues parlées en interaction permettent de mieux assurer dans le futur la collecte, l'archivage et le traitement scientifique, ainsi que la diffusion d'une très grande variété de corpus, dans des conditions légales satisfaisantes. De nouveaux outils informatiques ont été introduits dans le travail quotidien de la recherche, notamment pour le traitement automatisé et semi-automatisé d'anciens corpus et pour la construction des nouvelles données audio et vidéo. Les recherches sur les corpus ont aussi considérablement progressé, notamment en exploitant systématiquement, pour certains projets, les nouveaux outils de requête et les méthodologies analytiques qui leurs sont associées. Enfin, comme nous l'avons montré dans la dernière section consacrée aux exploitations scientifiques, le projet CLAPI permet d'envisager divers développements.

48 L'appareil méthodologique des approches interactionnistes peut être élargi au moyen des corpus CLAPI et des outils de requêtes intégrés dans la plateforme.

49 Un certain nombre d'arguments concernant les hypothèses de base de la linguistique interactionnelle, notamment le fonctionnement du système d'organisation des tours, peuvent être rendus plus robustes. Des recherches récemment entamées permettront ensuite de reprendre et d'approfondir des questions plus complexes concernant des clusters de phénomènes multimodaux – en incluant par exemple la prosodie et les gestes –, de structures séquentielles ou de l'infrastructure des tours.

50 Il sera possible de faciliter ainsi un rapprochement innovant entre les méthodes qualitatives systématiques et de plus en plus informatisées de la linguistique interactionnelle et des méthodes quantitatives appliquées jusqu'à présent essentiellement sur l'écrit.

51 La plateforme *Corpus de langues parlées en interaction CLAPI* compte parmi les rares bases de données interactionnistes en français. La quantité et la diversité de ces corpus comme les outils intégrés à la plateforme faciliteront les analyses qualitatives traditionnelles en linguistique interactionnelle, ainsi que des explorations dans le domaine de la linguistique de corpus et d'une nouvelle linguistique diachronique tenant compte des différentes dimensions de l'usage du français parlé en situation.

BIBLIOGRAPHIE

BALTHASAR, L., MONDADA, L. (2003) : Traitements qualitatifs et quantitatifs de corpus oraux : enjeux méthodologiques issus de la linguistique interactionnelle, *Colloque 36e Rencontre SLE « Linguistique et corpus »,* Lyon.

BANGE, P. (dir.), (1987) : *L'analyse des interactions verbales. La dame de Caluire. Une consultation*, Berne, Peter Lang.

BERT, M. et PLANTIN, C. (2003) : La base de données Corpus de Langues Parlées en Interaction (CLAPI), *Colloque 36e Rencontre SLE « Linguistique et corpus »,* Lyon.

BIBER, D., CONRAD, S., REPPEN, R. (1998) : *Corpus Linguistics. Investigating the language structure and use*, Cambridge, Cambridge University Press.

CASTEL, R., COSNIER, J., JOSEPH, I., QUÉRÉ, L. *et al.* (1990) : *Le parler frais de Erwin Goffman*, Paris, Minuit.

COSNIER, J., KERBRAT-ORECCHIONI, C. (dir.), (1987) : *Décrire la conversation*, Lyon, Presses universitaires de Lyon.

CRYSTAL, D. (1991) : *A Dictionary of Linguistics and Phonetics*, 3e édition, Oxford, Blackwell.

HEIDEN, S., BALTHASAR, L. (2003) : Outils de mise en place de corpus multimédiaux : le cas de la linguistique interactionnelle, *2nd Freiburg Workshop on Romance Corpus Linguistics : Corpora and Historical Linguistics*.

ICOR (BALTHASAR, L., BERT, M., BRUXELLES, S., ETIENNE, C., HEIDEN, S., MONDADA, L., PLANTIN, C., TRAVERSO, V.), (en préparation) : *Éléments de linguistique sur corpus de langue parlée en interaction*.

KENNEDY, G. (1998) : *An Introduction to Corpus Linguistics*, London, Longman.

KERBRAT-ORECCHIONI, C. (1990) : *Les interactions verbales*, Paris, Armand Colin.

MARTINS-BALTAR, M. et al. (1989) : *Entretiens, Transcription d'un corpus oral, Cahiers du français des années quatre-vingts, Hors-Série*.

MONDADA, L. (2001) : Pour une linguistique interactionnelle, in *Marges Linguistiques*, 1, 1-21.

MONDADA, L. (2004) : Temporalité, séquentialité et multimodalité au fondement de l'organisation de l'interaction : le pointage comme pratique de prise du tour, *Cahiers de linguistique française*, 26, 169-192.

MONDADA, L. (à paraître) : L'analyse de corpus dans la perspective de la linguistique interactionnelle : des analyses de cas singuliers aux analyses de collections, A. Condamine (dir.), *Sémantique et corpus*, Paris, Hermès.

PLANTIN, C., DOURY, M., TRAVERSO, V. (2000) : *Les émotions dans les interactions communicatives*, Lyon, PUL.

SACKS, H., SCHEGLOFF, E.A., JEFFERSON, G., (1974) : A Simplest Systematics for the Organization of Turn-taking in Conversation, *Language*, 50, 696-735.

SCHEGLOFF, E.A., (1992) : On talk and its institutional occasions, in P. Drew et J. Heritage, (dir.), Talk at Work, Cambridge, CUP, 101-134.

SCHEGLOFF, E.A., (1993) : Reflections on Quantification in the Study of Conversation, in *Research on Language and Social Interaction*, 26 (1), 99-128.

TIBERGHIEN, A., Jossem, E. L., Barojas, J. (dir.), (1998) : *Connecting Research in Physics Education with Teacher Education*, ouvrage en ligne, ICPE Books.

TRAVERSO, V. (1996) : *Conversation familière*, Lyon, PUL.

TRAVERSO, V. (1999), Négociation et argumentation dans la conversation familière, in *Escritos*, 31, Mexique, 51-89.

TRAVERSO, V. (à paraître), Malentendus, quiétude et inquiétude interprétatives dans la conversation familière, in M. Laforest (dir.), *Le malentendu : dire, mésentendre, mésinterpréter*, Québec, Nota Bene Éditeur.

TRAVERSO, V., Bruxelles, V. (2002) : Les corpus de langue parlée en interaction au GRIC, in C. Pusch et W. Raible (dir.), *Romanische Korpuslinguistik, Romance Corpus Linguistics*, Tübingen, Gunter Narr, 59-70.

Liste des corpus cités

Conversation familière, TRAVERSO, V. (1996), CLAPI, ICAR, Lyon2 – CNRS – ENS-LSH.

Conversations Pilat, MARTIN, J.-B. (2002), CLAPI, Institut Gardette/ICAR, Lyon2 – CNRS – ENS-LSH.

FRA 80, MARTINS-BALTAR, M., MOCHET, M.-A. (1989), CLAPI, ICAR, Lyon2 – CNRS – ENS-LSH.

Mode, Cosnier, J. (1987), CLAPI, ICAR, Lyon2 – CNRS – ENS-LSH.

TP Physique & Chimie, TIBERGHIEN, A., Le MARECHAL, J.-F. (1992, 2003), CLAPI, ICAR, Lyon2 – CNRS – ENS-LSH.

Conversations ordinaires, BALTHASAR, L., MONDADA, L. (2003-...), CLAPI, ICAR, Lyon2 – CNRS – ENS-LSH.

NOTES

1. Au niveau international, voir les fonds du *Linguistic Data Consortium* aux États-Unis (http://www.ldc.upenn.edu/), le *British National Corpus* en Angleterre (http://www.hcu.ox.ac.uk/bnc), les corpus de l'*Institut für deutsche Sprache* (http://www.ids-mannheim.de/), ceux de la *Real Academia Española* (http://www.rae.es), etc. Pour la situation francophone, voir par exemple les corpus *PFC* (http://infolang.u-paris10.fr/pfc/), *ELICOP* (http://bach.arts.kuleuven.ac.be/lancom/), *Valibel* (http://valibel.fltr.ucl.ac.be/corpus.htm), ainsi que le corpus de l'équipe DELIC (http://www.up.univ-mrs.fr/delic/rsfp/rsfp18.html).

2. Le développement de la plateforme CLAPI est piloté par le groupe de travail transversal ICOR de l'UMR ICAR : Ch. Plantin (direction), V. Traverso, L. Mondada, S. Bruxelles, C. Étienne, S. Heiden, M. Bert, L. Balthasar. Les travaux réalisés par ce groupe dans le cadre du développement de CLAPI font l'objet d'un ouvrage collectif, voir ICOR (en préparation).

3. Cf. Traverso & Bruxelles (2002), Bert & Plantin (2003), Balthasar & Mondada (2003), ICOR (en préparation). Pour une première version de la base CLAPI, voir : http://corpusgric.univ-lyon2.fr:8000/corpus/; un ensemble d'outils externes et internes sont accessibles sur http://weblex.ens-lsh.fr/projects/xitools/.

4. Cf. Bange (1987), Cosnier & Kerbrat-Orecchioni (1987), Castel et al. (1990), Kerbrat-Orecchioni (1990), Tiberghien et al. (1998), Plantin et al. (2000), Mondada (2001), ainsi que Mondada & Traverso, dans ce volume.

5. Mise à part la linguistique interactionnelle : socio- et ethnolinguistique, psycholinguistique, géolinguistique, linguistique de corpus ; en plus des sciences du langage, les sciences de l'éducation, les sciences cognitives, la psychologie, l'éthologie, etc.

6. *Laboratoire d'éthologie des communications*, direction Jacques Cosnier, puis *Groupe de recherche sur les interactions communicatives*, direction Catherine Kerbrat-Orecchioni, puis Christian Plantin.

7. Projet NOMEX-CLAPI réalisé en collaboration avec les laboratoires *Équipe de recherche en ingénierie des connaissances* (ERIC), Université Lumière Lyon 2 (http://eric.univ-lyon2.fr/), *Réseaux, information, multimédia* (RIM), École nationale supérieure des Mines de Saint-Étienne (http://rim.emse.fr/) ; en ce qui concerne le projet, voir aussi http://gric.univ-lyon2.fr/projets/nomex-clapi/index.html.

8. En ce qui concerne ces questions, des projets de recherches indépendants sont actuellement lancés ou en cours sous la direction de L. Mondada, V. Traverso et C. Plantin.

9. Pour la définition de *corpus* dans ce domaine, voir Crystal (1991), Biber *et al.* (1998), Kennedy (1998). Voir également la définition suivante : « Corpus : [...] a finite collection of machine-readable texts, sampled to be maximally representative of a language variety » (http://donelaitis.vdu.lt/publikacijos/SDoCL1.htm#Corpora).

10. La constitution des corpus CLAPI est donc souvent déterminée par des contraintes propres au terrain, par les dispositifs d'enregistrement, ainsi que par les techniques de transcription et d'actualisation des données pour l'analyse (voir ICOR, en préparation).

11. Pour le corpus FRA 80, voir Martins-Baltar et al. (1989).

12. Ces travaux sont dirigés par Andrée Tiberghien, voir par exemple le corpus « TP Physique & chimie ».
13. En ce qui concerne Dublin Core et ISLE, voir le site Internet http://dublincore.org/et http://www.mpi.nl/world/ISLE/ en ce qui concerne les descripteurs CLAPI, voir ICOR (à paraître).
14. Pour la TEI, voir par exemple le site http://www.tei-c.org/, pour un aperçu plus général, voir aussi http://www.ldc.upenn.edu/annotation/.
15. En ce qui concerne ces logiciels, voir les sites Internet http://childes.psy.cmu.edu/clan/et http://www.fon.hum.uva.nl/praat/.
16. Outre le format unifié ICOR, le moteur de requête de la plateforme CLAPI permet aussi l'affichage des transcriptions selon leur convention d'origine. En ce qui concerne la Convention ICOR, voir http://weblex.ens-lsh.fr/projects/xitools/format_xi/Xi. htm, ainsi que ICOR (à paraître).
17. Pour plus d'information concernant les formats utilisés, voir http://weblex.ens-lsh.fr/projects/xitools/format_xi/Xi. htm.
18. Centre d'études sur la coopération juridique internationale, collaboration dans le cadre du PSI Patrimoine et archivage documentaire. Les recommandations sont à paraître sur le site ICOR.
19. Une même personne peut apparaître sous plusieurs profils (ex. responsable d'un corpus et contractuel pour un autre corpus).
20. Voir par exemple Mondada & Traverso, dans ce volume.
21. Pour NXT Search, voir : <http://www.ims.uni-stuttgart.de/projekte/nite/> pour CLAPI QUERY, voir : <http://icar.univ-lyon2.fr/projets/requete_oral/index. html>
22.« Keyword-in-context ».

AUTEURS

LUKAS BALTHASAR

CNRS-ICAR

MICHEL BERT

Université Lyon2 / Université Catholique de Lyon – ICAR pour ICOR