CORPUS

Corpus

15 | 2016
Corpus de français parlé et français parlé des corpus

CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes

CLAPI, a multimodal database for talk in interaction: contributions and dilemmas

H. Baldauf-Quilliatre, I. Colón de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard et V. Traverso

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2991
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 octobre 2016
ISSN : 1638-9808

Référence électronique
H. Baldauf-Quilliatre, I. Colón de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard et V. Traverso, « CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes », *Corpus* [En ligne], 15 | 2016, mis en ligne le 15 janvier 2017, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2991

CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes

CLAPI, a multimodal database for talk in interaction: contributions and dilemmas

H. Baldauf-Quilliatre, I. Colón de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard et V. Traverso

1 Il est intéressant et possible, aujourd'hui, de mettre en perspective l'évolution des bases de données de langues parlées en France au cours des trente dernières années. Dans cet article, nous présentons le développement de la base de données CLAPI dans ce cadre. Nous détaillons les deux composantes de CLAPI, l'archive de corpus de langue parlée en interaction, audio et vidéo, enregistrés dans des situations sociales naturelles variées, et la plateforme d'outils. Nous montrons aussi comment la base peut être utilisée pour des études de linguistique interactionnelle à travers l'étude de « oh là là » et des usages de « trop » dans des contextes variés. Au cours de cette présentation, nous formulons quelques-uns des dilemmes auxquels nous sommes aujourd'hui confrontés dans les relations entre la poursuite des recherches sur des corpus variés (et parfois sensibles) et les exigences des bases de données ouvertes.

1. La base CLAPI et son contexte

2 La base de données CLAPI, Corpus de LAngue Parlée en Interaction a été lancée, à la fin des années 90, pour archiver et préserver les corpus qui étaient régulièrement faits dans le cadre des recherches sur l'interaction au laboratoire ICAR. Dès l'origine (1998-1999), la base a été pensée avec un triple objectif, qu'elle conserve toujours aujourd'hui (voir Bruxelles & Traverso, 2003). Elle s'est transformée au fil du temps pour devenir une plateforme outillée.

1.1 Objectifs de la base CLAPI

1.1.1 Une dimension « Patrimoine »

3 Sur ce plan, le développement et l'évolution de la base de données CLAPI sont représentatifs de la situation générale à au moins deux niveaux.

4 D'une part sur le plan de la réalisation d'une banque de données sauvegardant et mettant à disposition les corpus existants. Cet objectif a impliqué un important travail de recensement et de localisation des données, du fait qu'elles n'étaient jusque-là pas centralisées ni rendues disponibles à la fin d'une recherche. Les choses ont bien changé, depuis, des routines se sont mises en place, et la base héberge les corpus qui sont régulièrement réalisés, selon des standards qui ont été élaborés au cours du temps (voir ci-dessous). Ce processus est symptomatique des évolutions qui ont eu lieu au cours de la vingtaine d'années écoulée depuis le début de la conception de la base CLAPI. L'importance accordée aux corpus dans le champ scientifique (comme en témoignent les programmes de l'ANR qui y ont été consacrés) s'est démultipliée. Parmi les conséquences de cette évolution : l'attention plus grande portée à la collecte des données primaires et à la confection des corpus (transcription, organisation, etc., voir le site CORINTE[1]), la mise en place progressive de standards dans les manières de faire non seulement en informatique, avec le développement de la TEI au niveau international, par exemple, mais dans toutes les procédures conduisant à la réalisation des corpus (filmage, numérisation, transcription, etc.). L'archive de la base CLAPI conserve des traces de ce cheminement, avec des corpus historiques, et des corpus récents réalisés selon ces nouveaux standards.

5 D'autre part, les données hébergées dans CLAPI illustrent une très importante partie de l'histoire et du développement du champ d'analyse de l'interaction en France (voir Traverso, 2012b, Traverso *et al.*, 2012). Sont ainsi hébergés des corpus qui ont été réalisés par des chercheurs comme Bange, de Gaulmyn, Cosnier, Kerbrat-Orecchioni, Plantin, Bruxelles, Traverso, Grosjean, Mondada.

6 La constitution de l'importante archive de CLAPI (environ 600 heures) a impliqué un conséquent travail de sélection (selon des critères de qualité et juridiques) et d'organisation des données, comme la définition des entités « corpus », « interactions », « fonds », l'organisation des données primaires et des données secondaires, etc. (voir Balthasar & Bert, 2005). Sur le plan technique, cette réalisation a nécessité un important travail de numérisation (avec les choix techniques que cela entraîne) pour les enregistrements audio ou vidéo, qui existaient sur des supports extrêmement variés, aussi bien que pour les documents papier (données secondaires).

7 Un des problèmes majeurs qui s'est posé dans cette période concerne l'hétérogénéité des transcriptions, qui tenait à différents facteurs : l'utilisation de différentes conventions de transcription, la transcription partielle de certains phénomènes, de certains passages, les différents niveaux de granularité attestés, l'utilisation de différents logiciels de transcription (principalement CLAN, Praat et ELAN). La solution retenue conserve la transcription d'origine sans retranscription, dans le respect du travail effectué par le transcripteur, mais opère des modifications mineures qui sont consignées dans une version de la transcription « adaptée clapi », afin de résoudre des problèmes techniques comme l'utilisation d'un même signe pour des annotations différentes. Une procédure informatique transforme les annotations en balisages XML

qui sont utilisés par les outils de CLAPI pour traiter toutes les transcriptions quelle que soit leur convention, leur niveau de granularité ou leur format d'origine. Notre solution repose pour cela sur un processus qualité semi-automatique dans lequel l'équipe médiathèque intervient pour identifier et vérifier la convention fournie par le responsable puis détecter et corriger les anomalies. Ceci garantit la qualité des transcriptions présentes à ce jour dans la base, même si le volume actuel ne permet pas de corriger toutes les erreurs. Le responsable de corpus valide le choix des métadonnées et l'affichage des transcriptions avant que l'ensemble soit rendu disponible dans CLAPI.

1.1.2 Une dimension « Partage »

8 Sur ce plan, CLAPI entend faciliter la réalisation de recherches dans le domaine de l'interaction ou d'autres approches en linguistique en permettant aux chercheurs d'accéder à des données « toutes faites ». La mise à disposition des corpus s'accompagne :

– des descripteurs (75 métadonnées) ;
– du signal audio ou vidéo : en totalité, parfois uniquement l'audio pour des raisons de droit, et d'autres fois seulement des extraits ;
– des transcriptions : une transcription selon les principes de l'analyse conversationnelle à partir de laquelle on peut générer une transcription orthographique pour d'autres usages, dans différents formats ;
– d'un ensemble d'outils d'analyse et de requête.

1.1.3 Une dimension « Recherche »

9 La réalisation de la base CLAPI et la mise à disposition des données ont été pensées pour soutenir les analyses interactionnelles, qu'il s'agisse d'étudier la langue dans ses usages en interaction, ou plus conformément aux exigences de l'analyse conversationnelle, les configurations multi-ressources multimodales que les participants mettent en place dans leurs échanges. CLAPI, dans sa dimension de banque de données (archive) constitue un grand corpus permettant d'avancer sur la recherche des récurrences dans les organisations interactionnelles et, à partir de là, de constituer des collections (manière d'articuler le qualitatif au quantitatif).

1.2 CLAPI aujourd'hui

10 Dès sa conception, la base de données a présenté un certain nombre de caractéristiques qui marquent encore aujourd'hui sa spécificité parmi les bases existantes. C'est une base consacrée à la parole en interaction et non simplement au français parlé. Ceci conduit à accorder une attention très spécifique à la situation sociale dans laquelle les données sont collectées, ce qui a également pour conséquence :

– la très grande variété de situations sociales représentées dans la base (réunions de travail dans différents cadres, interactions de service, interactions en site commercial, visites privées, repas familiaux et amicaux, visites guidées, consultations médicales, appels téléphoniques privés et professionnels, situations de classe : travaux pratiques, conversations en ligne, etc.) ;

– le fait que les données hébergées dans la base sont très majoritairement des « données naturelles ». On désigne par cette expression le fait que les données ne sont pas produites pour le chercheur ni dans une situation construite par le chercheur (Potter, 2006). La plupart des données de CLAPI sont des enregistrements d'interactions se déroulant dans leur milieu habituel et pour leurs raisons habituelles propres aux participants. On peut souligner que cette « naturalité » distingue les données de CLAPI de la plupart des données orales que l'on trouve le plus souvent dans les bases de données, et qui sont provoquées ou obtenues par élicitation (p. ex. : des entretiens[2]). La différence entre les deux est particulièrement signifiante pour les situations de travail. Parler de naturalité n'implique pas que l'on considère que le protocole d'enregistrement n'a aucun impact sur les comportements des participants (voir Colón de Carvajal *et al.*, à paraître, Laurier & Philo, 2006). Toutefois, et malgré la présence de la caméra et son impact, les données naturelles sont irremplaçables pour étudier les processus interactionnels en situation.

11 Les données hébergées dans CLAPI, originellement audio, sont aujourd'hui de plus en plus souvent vidéo.

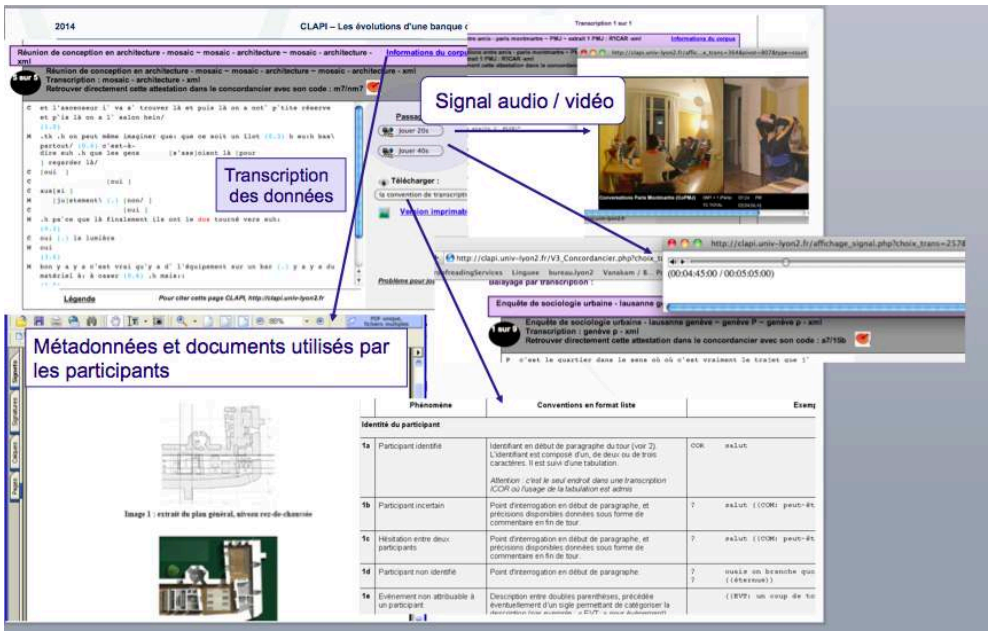
12 Outre les données qui ont été collectées par les chercheurs du laboratoire ICAR, CLAPI héberge des données d'interactions confectionnées par d'autres équipes de recherche, et dont le processus d'intégration dans la base (métadonnées, transcriptions, accès, etc.) est discuté avec les auteurs. Sont actuellement hébergés : les Cahiers du français des années 80 (M.-A. Mochet), un Fonds Bielefeld (E. Gülich), le Corpus Grenouille (H. Jisa), le Corpus Étudiants (M. Savelli), le Corpus Entretiens avec des jeunes écoliers (J.-M. Colletta), etc.

1.2.1 L'organisation

13 La base de données CLAPI est une base de données multimédia au sens fort. L'organisation des données qu'elle contient est conçue de telle sorte que, pour chaque corpus (qui correspond à un seul enregistrement dans le cas le plus simple), il est possible d'accéder à l'ensemble des éléments documentant ce corpus : le signal audio et vidéo par streaming ou téléchargement, la transcription des données, les conventions de transcriptions, les autres données primaires (documents récupérés sur le terrain), et les métadonnées (voir Figure 1). L'ensemble de ces éléments est accessible aussi bien à partir de la fonction « feuilleter les corpus » qu'à partir des résultats d'une requête effectuée à l'aide d'un des outils de la plateforme.

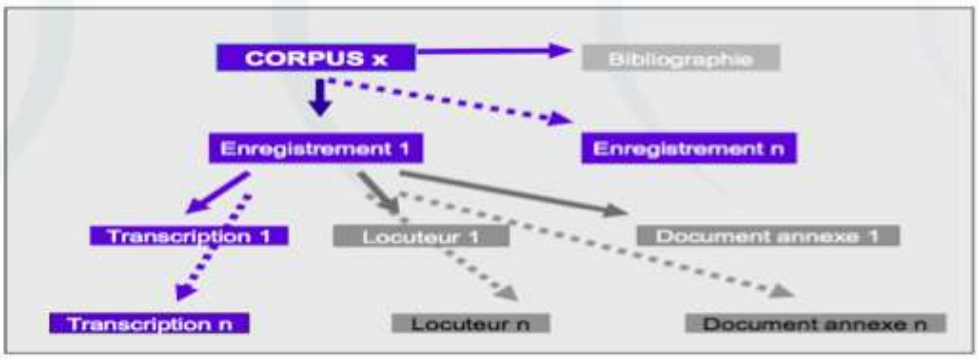
14 Cas simple : un corpus, une situation, une interaction.

Figure 1. Données et métadonnées



15 Les cas plus complexes sont ceux dans lesquels le corpus comprend plusieurs enregistrements, ou comporte plusieurs transcriptions (le choix ayant été fait de proposer différentes transcriptions dans différents formats pour un même enregistrement). L'architecture est alors la suivante (Figure 2) :

Figure 2. Architecture



1.2.2 Les outils

16 L'analyse des interactions repose sur l'étude fine de phénomènes dans une approche résolument qualitative ; les outils de CLAPI permettent d'introduire une dimension quantitative (décompte de phénomènes) qui peut conduire à formuler de nouvelles hypothèses de recherche et sont ainsi une aide pour construire un objet d'étude complexe.

17 La palette d'outils mise à la disposition des usagers est une aide pour approcher un phénomène donné, qu'ils pourront aborder à partir des repérages automatiques, puis en retournant systématiquement au détail des attestations transcrites, à l'écoute ou à la visualisation de l'enregistrement, voire aux métadonnées (*cf.* ci-dessus).

18 L'interface est disponible, depuis 2013, en langue anglaise pour permettre aux chercheurs étrangers de disposer de collections de données en français qu'ils pourront néanmoins sélectionner dans leur langue.

19 Le développement de ces outils a nécessité un travail préalable sur l'orthographe utilisée dans les transcriptions. Les transcriptions originales sont en effet le plus souvent réalisées en orthographe adaptée (voir les conventions ICOR[3]), c'est-à-dire cherchant à reproduire à l'écrit certains aspects de la prononciation (par exemple, *`fin* et non *enfin*, *b`jour* et non *bonjour*), un outil a été développé pour reconstruire (et permettre de travailler sur) la « forme étendue » afin de retrouver l'ensemble des attestations. Le même outil permet de générer une transcription orthographique standard des corpus.

20 Les autres outils développés sont les suivants :

– Un outil de concordance permet de retrouver un token dans toute la base ou dans un sous-ensemble de corpus (à partir de filtres sur la nature du signal, audio ou vidéo, et sur le nombre de locuteurs) ;

– Une série d'outils automatiques produisent des résultats quantitatifs à partir d'un point d'entrée qui peut être un mot, une transcription ou un phénomène interactionnel. Ils comprennent :

- le lexique d'une transcription, par fréquence ou par ordre alphabétique ;
- les co-occurrences d'un mot dans toute la base ou dans un sous-ensemble de corpus, c'est-à-dire les mots les plus fréquents dans le voisinage gauche ou droit du mot cible ;
- les co-occurrences d'un phénomène interactionnel. Les phénomènes traités sont les chevauchements (avec la distinction chevauchant/chevauché), les pauses (courtes/longues), l'emplacement dans le tour de parole, les tours courts ;
- les contextes d'emploi d'un mot : les emplois les plus fréquents du mot cible en fonction de sa position dans le tour (seul, en première position dans le tour, en début de tour, en dernière position dans le tour, en fin de tour, dans les tours courts), de sa production en chevauchement ou non (début de segment chevauchant ou chevauché), de sa localisation par rapport aux pauses, etc. ;
- les répétitions dans une transcription, qu'il s'agisse d'auto-répétitions ou d'hétéro-répétitions. L'outil identifie les segments les plus répétés, par fréquence et par taille, dans la transcription complète ou par locuteur.

- Un outil de requêtes multicritères associe le lexique, les caractéristiques interactionnelles et les métadonnées. Il permet au chercheur de définir lui-même son objet d'étude qui peut correspondre à :

- une expression composée d'une suite de mots, à une certaine distance, dans le même tour ou dans une suite de tours de parole ;
- à une certaine position du tour de parole (n tokens du début ou de la fin du tour), dans des tours de parole d'une longueur donnée (plus de/moins de n tokens) ;
- avant ou après une pause, courte et/ou longue ;
- en début de segment chevauchant/chevauché ;
- dans des interactions sélectionnées à partir des critères : audio/vidéo ; nombre de locuteurs ; type d'activité situations ; locuteurs natifs ou non natifs ; sexe ; tranche d'âge ; ou par un locuteur donné (si un sous-ensemble de corpus a été sélectionné).

21 Les résultats des requêtes effectuées avec ces outils permettent de retourner à l'ensemble des informations présentées au paragraphe 1) ci-dessus. Ils donnent en outre la possibilité d'accéder pour chaque extrait, à une version « détaillée »

(transcription fine) ou « simplifiée » (transcription orthographique), ainsi qu'à une version imprimable qui permet de copier/coller un passage dans un autre document.

1.2.3 Quelques données chiffrées

22 CLAPI comprend, à ce jour (mars 2015), 65 corpus correspondant à 370 situations interactionnelles, soit 225 heures de données et 650 transcriptions, les transcriptions alignées étant disponibles dans plusieurs formats. Parmi ces données, 45 heures sont téléchargeables sans condition d'accès, et 65 heures, soit 150 situations, sont requêtables par les outils décrits ci-dessus.

23 Les consultations représentent environ 10 000 accès par mois, en excluant la page d'accueil ou les requêtes qui ne sont pas formulées jusqu'au bout par l'utilisateur. On peut détailler parmi ces accès : 30 % de consultation des métadonnées ; 30 % de téléchargement des enregistrements et des transcriptions mais aussi des conventions de transcription ; 20 % d'utilisation des outils et 10 % de streaming des enregistrements. Les outils les plus utilisés restent les concordances (30 %) et les requêtes multicritères (30 %), les outils automatiques se partagent les 40 % restant, sans préférence marquée pour l'un d'entre eux.

1.2.4 Les autres sites en relation directe avec CLAPI

24 La base met à disposition un espace de travail (de type « bac à sable ») pour les corpus en cours d'exploitation (projets, thèses, etc.) qui donne accès à l'ensemble des outils d'analyse et de requête de la base tout en nécessitant un jeu restreint de descripteurs. L'enjeu est aussi de favoriser le dépôt depuis cet espace vers la banque de données, à la fin des projets ou des thèses.

25 CLAPI est associé avec le site CORINTE (CORpus d'INTEractions)[4] qui est dédié à la méthodologie et aux aspects analytiques de la linguistique interactionnelle, explicitant toute la chaîne de production des corpus, les questions juridiques et les principes d'analyse, et mettant différents documents à la disposition des utilisateurs (p. ex. : autorisations, consentement éclairé, etc.).

26 La base est également associée avec le site CORVIS (CORpus de VIdéos Situées)[5] qui recense les usages de la vidéo en sciences humaines et sociales, en vue de la constitution de corpus pour l'étude des pratiques sociales, culturelles, linguistiques dans leurs contextes ordinaires, professionnels et institutionnels. Le site rassemble de nombreuses informations pour la réalisation et le traitement des vidéos.

2. Évolutions majeures

27 La base de données a évolué sur tous les plans au fil du temps. Nous ne reprenons que les éléments majeurs, qui sont aussi l'occasion d'évoquer les problèmes et les dilemmes qui se posent.

2.1 Alimentation et enrichissement de la base : les nouveaux corpus vidéo

28 Comme nous l'avons dit ci-dessus, les nouvelles données sont réalisées à partir de standards qui ont été établis au fil du temps (*cf.* les sites CORINTE et CORVIS). La chaîne de production des corpus est intégrée dans le cursus de formation des étudiants de sciences du langage : réalisation des terrains, filmage (prise de vue, conception), transcription outillée (CLAN, Praat, Transicor, ELAN). Ceci permet à la base CLAPI de mettre à la disposition des chercheurs des enregistrements vidéo, le plus souvent multivue, d'excellente qualité, qui sont propices à l'étude de phénomènes interactionnels multimodaux les plus divers (voir Mondada, 2006). Ces évolutions inestimables pour la recherche (en termes de variété de données, de qualité du signal vidéo et audio, et de démultiplication des phénomènes rendus étudiables parce qu'accessibles) ne vont pas sans poser des problèmes et nous confronter à des dilemmes.

29 Par exemple, le nombre des tâches liées à la mise en forme et à la mise en ligne d'un corpus augmente en parallèle. En plus des tâches de numérisation (*i. e.* le transfert du format natif de la caméra vers un format qui soit interopérable entre players et systèmes d'exploitation, et compressé sur ordinateur), apparaît celle de synchroniser les sources. En effet, pour favoriser une lecture complète des données multivue sur CLAPI, les différentes vues enregistrées (et les sources audio additionnelles, s'il y en a) sont synchronisées en une seule vidéo à l'aide de logiciels professionnels (FinalCut Pro). Ceci permet également au chercheur d'activer à l'écoute une source audio plutôt qu'une autre (de meilleure qualité ou de meilleur volume sonore). Cette multiplication des sources (audio et vidéo) pose des problèmes liés au poids et au volume des données à archiver. Tout l'ensemble du processus pour une valorisation optimale des données nécessite au final de plus en plus de tâches, de plus en plus techniques (transfert, compression et synchronisation des données) et, en conséquence, une augmentation en effectif humain et en recherche continue de financement.

30 Un des aspects non résolus de ce dilemme concerne les dimensions multimodales. Les travaux menés dans l'équipe LIS du laboratoire ICAR intègrent de façon aujourd'hui systématique la multimodalité (*cf.*, entre autres, Mondada, 2006, 2007, 2012 ; Groupe ICOR, 2014 ; Traverso 2011, 2012a, 2014 ; Ticca & Traverso, à paraître ; Baldauf-Quilliatre, 2014a et b ; Colón de Carvajal, 2013). Les analyses réalisées sont possibles grâce à la qualité des données collectées. Ces données sont hébergées dans la base (ou dans l'espace de travail privé de CLAPI), mais elles ne sont pas annotées multimodalement. La réalisation d'une analyse multimodale implique de suivre la démarche présentée dans ICOR 2014, que l'on peut résumer ainsi :

> Parcours de la base, requête, résultats de la requête -> établissement de la collection (par sélection) et classement
> Retour aux données (signal) -> nouveau travail sur la transcription en fonction des besoins de
> la recherche (granularité, annotations multimodales pertinentes pour l'analyse) -> analyse multimodale

31 Les questions qui se posent concernent d'une part la pertinence de poursuivre la réalisation d'aussi nombreux nouveaux corpus (notamment dans le cadre de la formation), sachant qu'il n'est pas possible de les traiter, transcrire et intégrer (ni en totalité ni rapidement) dans la base de données. L'autre question est celle de la

pertinence de réaliser une annotation multimodale des données dans CLAPI, qui ne pourrait de toute façon que porter sur un très petit nombre de données (comparativement à ce que la base met à disposition), et sur un ensemble très restreint de phénomènes par rapport à ceux que l'analyse interactionnelle fait jouer. Ces questions continuent à être en discussion dans l'équipe de gestion de CLAPI.

2.2 Les dimensions juridiques

32 Il y a une vingtaine d'années, le recueil de données audio était réalisé le plus souvent sans précaution particulière. Depuis, l'obtention du consentement des personnes enregistrées est devenue une étape indispensable avant toute prise de données. Ce changement s'explique notamment par l'usage de la vidéo, qui fait apparaître les visages en plus des voix, et par le développement des bases de données de corpus en ligne, qui favorise la diffusion de ces images. L'enregistrement, l'exploitation et la diffusion des données audiovisuelles illustrant des situations d'interaction ordinaires de la vie quotidienne posent des questions de droit des personnes enregistrées (droit à la vie privée et droit à l'image), et des questions d'éthique relatives à la diffusion des données enregistrées. Après la collaboration du groupe ICOR aux réflexions collectives qui ont mené à la publication du guide des bonnes pratiques (Baude, éd., 2006), le travail effectué en 2009 avec les services juridiques du CNRS a permis de rendre CLAPI conforme aux évolutions de la réglementation en matière de protection des données dites « à caractère personnel ».

33 Un des changements qui en découle est la mise en place de Conditions générales d'utilisation (CGU), explicitant la restriction de l'utilisation des données à des fins de recherche et d'enseignement, qui doivent être acceptées (de manière électronique) pour toute consultation des corpus.

34 La question de la diffusion des données de la recherche est plus que jamais au cœur des préoccupations de la communauté des SHS. Les initiatives locales sont nombreuses autour de ces questions pour tenter d'en définir les contours et de trouver des solutions. Mais il importe qu'elles soient traitées à l'échelle nationale pour aboutir à des directives communes et à l'harmonisation des pratiques. C'est ce que l'on attend des travaux du consortium IRCOM ou du réseau des MSH, auxquels le groupe ICOR contribue, ainsi que des infrastructures en réseau, comme ORTOLANG, qui proposent des services mutualisés d'archive pérenne de données et de diffusion à grande échelle.

35 Globalement, la tendance actuelle est à l'ouverture de plus en plus importante des données de la recherche. Cette évolution est une conséquence logique de la mise en place des bases de données, tout à fait positive sur le plan du rayonnement de la recherche et de la qualité des données. Elle pose en retour quelques problèmes, par exemple celui de décider si l'on doit continuer à confectionner des corpus auxquels l'accès sera toujours restreint (p. ex. : corpus en milieu médical). C'est à nouveau tout l'équilibre entre force de travail, coût, reconnaissance et diffusion qui se trouve posé, des positions trop radicales en la matière risquant d'avoir un effet appauvrissant sur la diversité des domaines étudiés.

2.3 Les interopérabilités

36 Sur ce plan également, le panorama n'a cessé d'évoluer au cours des quinze dernières années.

37 Un premier besoin d'interopérabilité bilatérale a émergé dans les projets comprenant plusieurs bases de données pour échanger les métadonnées et les transcriptions, voire accueillir les corpus dans les différentes bases afin de bénéficier d'une plus grande variété d'outils d'exploration ou de requêtes. CLAPI a ainsi développé une plateforme CLAPI-TALKBANK dédiée aux corpus d'Analyse conversationnelle de la TALKBANK[6] en anglais et en danois, basée sur le format XML de la TALKBANK. Ce type d'interopérabilité implique un suivi permanent pour s'assurer qu'un changement effectué dans une des bases ne fasse pas barrière à l'interopérabilité. Chacune des bases étant en évolution constante, cette solution ne peut pas être maintenue à moyen terme.

38 Pour éviter de multiplier des formats pivots voués assez vite à devenir obsolètes, CLAPI a proposé, dès 2006, un export de ses descripteurs et de ses transcriptions en format TEI[7]. Ce recours au format TEI a été exploité par la suite dans l'ANR franco-allemande CIEL-F[8] « Corpus international écologique de langue française ». Il a permis l'échange d'une collection de métadonnées entre les bases MOCA et CLAPI (les transcriptions sont en Praat), et une plateforme CLAPI-CIELF est en cours de finalisation proposant la palette d'outils de CLAPI pour explorer les corpus de CIEL-F.

39 Une réflexion plus générale a été initiée dans le groupe de travail « Interopérabilité » (coord. C. Étienne, ICAR, C. Parisse, Modyco), au sein de l'infrastructure de recherche IRCOM[9] dédiée à l'étude des Corpus oraux et multimodaux en partenariat avec l'équipex ORTOLANG[10]. Ce groupe participe aux discussions du groupe européen ISO-TEI pour proposer des évolutions dans la norme adaptées aux spécificités de l'oral. Ses objectifs sont de convenir d'un jeu raisonnable de métadonnées indispensables à tout travail de recherche ainsi que d'un format commun de transcriptions pour permettre aux chercheurs de travailler sur une plus grande quantité de données, quels que soient leur base d'origine, leur structure initiale et le format de leur transcription. Cette initiative a été enrichie par les besoins d'homogénéisation des corpus oraux du projet ANR ORFEO[11] « Outils et ressources pour le français écrit et oral ». Il est clair aujourd'hui que l'interopérabilité ne peut être traitée au sein d'un seul laboratoire et que c'est collectivement que l'on peut proposer des solutions s'adaptant à la variété des données du paysage de l'oral.

3. Quelques exemples de recherche

40 Nous présentons succinctement deux exemples de recherches qui ont été effectuées pour illustrer des usages possibles des outils proposés par CLAPI. Le premier exemple illustre le travail de mise en relation des données et des métadonnées pour l'analyse et le second, la façon dont la base de données peut permettre de travailler sur la multimodalité.

3.1 Trop : articulation données et métadonnées

41 À la suite du travail d'O. Daumeries dans un dossier de M[aster 2], nous avons repris l'analyse des usages de « trop », en posant qu'il existait un glissement de son sens

« originel » dénotant l'excès (« c'est trop haut pour que je l'attrape peux-tu m'aider ») jusqu'au sens de simple intensif « c'est trop beau ». Nous avons sélectionné les corpus en fonction de la langue des locuteurs, en excluant les interactions dans lesquelles intervenaient des locuteurs non natifs, et n'avons sélectionné que les corpus enregistrés en France (par l'examen de la liste des corpus, ou à travers les métadonnées).

42 Au total, nous avons retenu 24 corpus (*cf.* le tableau en annexe). L'étude des occurrences de « trop » fait apparaître d'emblée que la simple opposition « intensité » *vs* « excès » n'est pas suffisante pour la description. Nous relevons les emplois suivants.

3.1.1 Emploi « trop = excès »

43 On peut distinguer ici plusieurs sous-catégories.

- Les emplois « classiques » :

(1) CEC : ça a été un peu **trop** assimilé à mon avis à la loi Pasqua qui est plus la fermeture (Débat sur l'immigration)
(2) EF : est-ce que vous trouvez que par exemple à la télévision justement euh on parle **trop** de: enfin de mort (Cahiers du Français des années 80)
(3) Y : mais c`t après-midi là: tu vois j` su- j` suis descendu en ville t`t à l'heure en milieu d'après-midi (.) j'avais même chaud (.) j'étais euh: **trop** habillé tu vois (Conversations familières, Navye)

- Les emplois classiques qui correspondent à la négation ou à la remise en question de la notion d'excès :

(4) C5 : vous avez pas **trop** froid en vélo (Interactions pendant la tournée des facteurs)
(5) Ap6 : tout tout juste en espérant qu'il n'y a pas **trop** d` trafic hein ça va dev`nir la mauvaise heure hein non/ (Téléphone en entreprise)

- Cas de « de trop »

44 On trouve deux occurrences de « de trop »[12] dans la base de données :

(6) FA17 : bon il faut pas qu'il en fasse **de trop** non plus (Enquête de sociologie urbaine - paris marais)

3.1.2 « Trop » emploi intensif au sens de « très », « tellement », « beaucoup »

45 Le TLFI rappelle que ces emplois sont attestés depuis longtemps dans certains contextes :

- les formules de politesse, comme « vous êtes trop aimable, trop bon, etc. » ;
- dans des tours hypocoristiques, par exemple : « Ils se retiraient sur la pointe des pieds en murmurant que j'étais trop mignon, que c'était trop charmant » (Sartre, *Les Mots*, 1964, p. 119).
- et dans des phrases exprimant une appréciation subjective, exemple : « Ah ! non c'est trop drôle ! Ah ! ah ! ah ! » (Feydeau, *La Dame de chez Maxim*, 1914, II, 8, p. 48).

(7) ELI : ça m'a **trop** peinée (0.2) franchement (Repas Kiwi)
(8) FLO : c'est **trop** bon ça mh::\ (Repas Olives)

46 Sur cette base, la mise en parallèle des occurrences et des métadonnées nous permet de faire les observations suivantes.

3.1.3 Âges des locuteurs et époque d'enregistrement

47 L'époque d'enregistrement est indéniablement pertinente. Par exemple dans les corpus de conversations familières enregistrées entre 1985 et 1990, il n'y a qu'une seule occurrence de l'emploi de « trop » au sens de « très » :

(9) A : c'est vraiment **trop** drôle parce qu'y a un moment où Mozart est occupé/ (.)
alors y a quelqu'un qui dit he is busy (Conversations familières, Navye)

48 Et l'on peut noter qu'il s'agit d'un usage répertorié dans le TLFI. Dans les corpus correspondant aux mêmes situations qui ont été enregistrés en 2008 (Épinards, Kiwi, Olives), ces emplois sont largement supérieurs aux emplois au sens classique :

Figure 3. Fréquences d'emploi

Corpus	Année	durée min	nombre d'occ.	Trop Excès	Trop intensif
Repas Épinards	2008	31	8	2	6
Repas Kiwi	2008	150	88	2	86
Repas Olives	2008	29	20	2	18

49 D'une façon générale, on peut dire que cet usage semble donc occasionnel, jusqu'aux années 2000, puis devient plus conséquent.

50 Pourtant, la période d'enregistrement et l'âge des participants n'expliquent pas tout.

3.1.4 Genre interactionnel

51 L'autre élément essentiel est le genre interactionnel. Dans le corpus Session de jeux vidéo, enregistré en 2007 (dans le but d'étudier la langue des jeunes), dans lequel des adolescents jouent à un jeu de football, les occurrences « classiques » sont plus nombreuses que les occurrences comme simple intensif. C'est dû au fait que les participants commentent les tirs et les manières de jouer, avec une grande fréquence d'énoncés comme :

(10) j` vais trop vite
trop haut
ah putain trop court

3.1.5 Les « préfabriqués »

52 Une autre piste de réflexion est ouverte par cette première étude, c'est celle des « préfabriqués » (constructions toutes faites, *chunks*) (voir Gülich, 2008 ; Schmale, éd., 2013).

53 Nous obtenons 56 occurrences de « pas trop » dans les corpus sélectionnés, parmi lesquelles

- 10 « j'aime pas trop »

(11) la paella j'aime pas trop (Repas Olives)
(12) moi j` trouve ça fait un peu boyau\ ça mais moi j'aime pas trop (Réunion de conception en Architecture, Mosaic)

- 7 « savoir » + pas trop

(13) bon ben j'ai fichu mon b- pas mon beurre dessus puisque c'est du St Hubert ou **j` sais pas trop** quoi faut que j` m'entretienne la ligne (Interactions dans un commerce - magasin de retouches)

54 Un nombre important d'occurrences concerne des énoncés sans verbe réalisant des évaluations, dont les outils de CLAPI permettraient d'étudier l'emplacement séquentiel par rapport au tour précédent, notamment si ces « *assessments* » sont produits en chevauchement :

« trop bien » 31 occurrences
« trop beau » 8
« trop fort » 7

3.2 « Oh là là » une façon de travailler sur la multimodalité à partir de CLAPI

55 Le deuxième exemple illustre comment CLAPI peut permettre l'analyse de la multimodalité dans une approche qualitative de linguistique interactionnelle. L'étude poursuit l'investigation de différents marqueurs discursifs en interaction effectuée par le groupe ICOR (2007, 2008a, 2008b, 2009, 2010). Dans ce cadre, nous nous sommes intéressés à « Oh là là » (ICOR 2014). « Oh là là » est souvent décrit comme un exclamatif typiquement français, pouvant servir à marquer la surprise ou la consternation. Cette explication n'est pourtant pas suffisante comme l'a montré notre étude basée sur des analyses multimodales. CLAPI nous a permis de faire une collection de 67 occurrences dont 59 ont finalement été retenues, les 8 autres étaient prononcés par des locuteurs non-francophones ou insuffisamment audibles pour une analyse. Ces 59 exemples pouvaient être regroupés dans deux grandes catégories. La première catégorie regroupe les cas où « Oh là là » est utilisé (seul ou avec d'autres éléments langagiers) en tant que première ou deuxième partie d'une paire adjacente ou en tant que continueur. Dans ce cas, le marqueur participe à la co-construction de l'interaction. La deuxième catégorie regroupe les cas où ce n'est pas à ce niveau qu'intervient le marqueur : soit parce qu'il se trouve dans un tour long, narratif et introduit un discours rapporté, soit parce qu'il réfère à une activité / un événement extralinguistique.

56 Les documents vidéo disponibles dans CLAPI nous permettent de prendre en compte toute la dimension multimodale de l'interaction (agencement de l'espace, gestes, regards, position, manipulation d'objets etc.). Dans l'exemple suivant, extrait d'une interaction dans un tabac-presse, une cliente entre dans le magasin avec un journal qu'elle a pris à l'entrée sur un présentoir et non pas sur le distributeur. BEA, la vendeuse, lui fait une remarque (l.01-02).

(14)

```
   01 BEA     vous voulez pas prendre/ celle qu` était
   02         sur le ◊distri+bu◊teur/ °hein°
      cl64           ◊pose le journal sur le comptoir
      bea                   +G-> cl64
      cl64                     ◊G->entre>>
   03         (0.5)
   04 CL64    #°j'ai pas pensé°
      im      #Im1
   05         (0.3)
->    06 CL64    #◊oh: là là/ [hin hin
      cl64     ◊met sa main devant son visage >>
      im      #Im2
   07 BEA                  [ha HA
   08 CL64    #<((en riant)) c'est que MAINt`nant//&
      im      #Im3
   09 CL64    &que je l` v[ois:/]>
   10 BEA                 [mais c'est] pas <((en riant))
   11         gra:ve °allez-y°> ((rire))
   12 CL64    je l'ai cherché (je me suis dit) bon:\
   13         allez [...]
```

57 Cette remarque est une critique à l'égard de la cliente et pointe la non-conformité de la manière de faire qu'elle a adoptée. La cliente répond d'abord avec une excuse assez vague (« J'ai pas pensé », l. 04). La suite (« Oh là là hin hin », l. 06) pourrait être interprétée comme une expansion de cette excuse. L'analyse détaillée des gestes et des regards montre autre chose. Lorsque la cliente répond « J'ai pas pensé », elle regarde en direction de l'entrée (Figure 4a). Elle maintient ce regard pendant la pause de 0.3 sec. (l. 05), puis met ses mains devant son visage, fait un pas en arrière et s'exclame en riant « Oh là là hin hin » (l. 6, Figure 4b). La vendeuse rit à son tour. La cliente pointe alors vers le distributeur et explique qu'elle n'a pas pris le journal dessus car elle ne l'avait pas vu (l. 8-13, Figure 4c).

Figure 4. Langage paraverbal

Im. a. BEA regard à CL64, pendant que CL64 regarde vers l'entrée	Im. b. CL64 met sa main devant son visage avant de s'exclamer oh là là et de rire.	Im. c. Les deux participantes rient. CL64 pointe vers le distributeur et dit en riant « c'est maint`nant que je l` vois. »

58 *Oh là là* indique ici un changement d'état : entre la réponse « J'ai pas pensé » et « Oh là là » la cliente a aperçu le distributeur et compris son « erreur ». Regards, position et pointage sont des indices forts pour comprendre « Oh là là » comme change-of-state

token (Heritage, 1984) et introduisant l'explication, et non comme expansion de l'excuse.

59 La démarche pour ce type d'analyse à l'aide de la base CLAPI est celle que nous avons présentée dans la section 2.1 :

> Parcours de la base, requête, résultats de requête -> établissement de la collection (par sélection) et classement
> Retour aux données (signal) -> nouveau travail sur la transcription en fonction des besoins de la recherche (granularité, annotations multimodales pertinentes pour l'analyse) -> analyse multimodale

4. CLAPI pour l'enseignement

60 Depuis 2010, nous avons engagé une réflexion sur la façon dont la base de données pourrait être utilisée pour l'enseignement du français (FLE) ou de la linguistique française, l'idée étant de concevoir un volet de CLAPI dédié à l'enseignement qui serait alimenté en parallèle du volet recherche.

61 Nous avons collaboré avec plusieurs départements de français à l'étranger où des expériences d'utilisation des données de CLAPI ont été réalisées :

- en groupe classe, avec des étudiants de niveau A2 à B1, utilisation de corpus de CLAPI choisis par l'enseignante pour illustrer différents phénomènes interactionnels comme les routines, les assimilations, les émotions, les élisions (ICOR et E. Ravazzolo, Université de Trento, Italie)
- en groupe classe, avec des étudiants de niveau C1 futurs interprètes, utilisation d'extraits portant sur le désaccord avec des tests de compréhension et de reformulation (ICOR et N. Niemants, U. Forli et Maccerata, Italie)
- en entretien individuel auprès de 9 étudiants de niveau A2 à C1, utilisation d'extraits portant sur le désaccord avec des tests de compréhension, de reformulation, de détection de début et de fin de séquences et leur justification (ICOR et A. Thomas, J. Granfeldt, N. Bengtsson & C. Rocher-Hahlin, U. Lünd, Suède, dans le cadre du projet exploratoire Clapi-FPIE[13])

62 Dans tous les cas, l'enseignant(e) a donné des informations contextuelles et parfois lexicales, et constaté que les difficultés des élèves portaient davantage sur l'organisation de l'interaction, la co-construction de tours de parole par plusieurs locuteurs et la compréhension des tours brefs que sur le lexique. Les expériences ont également montré que le niveau de bruit (souvent considéré comme un frein à l'utilisation de données enregistrées dans des contextes sociaux naturels) n'est pas toujours problématique et constitue au contraire une aide à la compréhension.

63 Cet intérêt des enseignants pour l'usage de données orales naturelles en complément des données construites a conduit à ouvrir une collaboration avec des chercheurs engagés dans des directions similaires pour d'autres bases de données (PFC, S. Detey & I. Racine ; ESLO, M. Skrovec).

5. Conclusion

64 Le temps n'est pas si lointain où l'on se plaignait, à juste titre, de l'absence de corpus de français (parlé). La situation a fortement évolué au cours des vingt dernières années. Il ne semble plus guère possible de dire aujourd'hui qu'il n'y a pas de corpus disponibles

pour travailler, et les récriminations relatives à l'impossibilité d'accès (*cf.* l'expression si souvent entendue « du chercheur assis sur ses données ») n'ont certainement plus lieu d'être, que les données soient en ligne ou qu'elles soient accessibles après un contact avec les gestionnaires des bases de données. Le développement de ces bases a fait avancer les méthodologies et la recherche dans de nombreuses directions (sur le plan de la qualité des données, de la connaissance des attentes d'un public élargi, des exigences de standardisation pour permettre le partage, de la quantité de données à disposition ce qui entraîne un enrichissement des analyses, etc.). L'expérience de CLAPI que nous avons retracée ici est tout à fait représentative à cet égard.

65 Dans le panorama actuel, on voit combien les bases de données existantes sont complémentaires et présentent chacune ses spécificités. CLAPI est ainsi la seule à être spécifiquement dédiée à la langue parlée en interaction, ce qui la conduit à proposer une très grande variété de corpus vidéo enregistrés dans des situations sociales variées. Elle se caractérise tout autant par la riche panoplie d'outils qu'elle met à la disposition des chercheurs.

66 Se dessine par ailleurs, aujourd'hui, une très nette orientation vers la collaboration entre les chercheurs des différentes bases (dans les projets ANR ou dans les instances nationales) pour réfléchir à des solutions communes plutôt que de continuer à avancer en parallèle sans concertation, même si les objectifs et les contenus des bases restent différents sur bien des points.

67 La situation nous conduit également aux constats suivants :

– Il manque encore un très grand corpus de français. Comme le préconise le projet ORFEO, c'est en fédérant et organisant les bases et corpus existants que ce très grand corpus a des chances de se mettre à exister ;
– D'une façon plus générale, et à toutes sortes de niveaux, on peut dire que l'effort qu'il a fallu faire au cours des années 1990 pour commencer à réunir les forces sur la centralisation des corpus, leur identification, leur conservation, etc. et pour lancer les projets de bases de données (*cf.* Bruxelles & Traverso, 2003) se poursuit actuellement avec l'organisation des collaborations entre bases de données.

68 En conclusion, il nous semble important de rappeler que ces évolutions créent également des problèmes voire des dilemmes. Les bases de données, les exigences de standardisation, la lourdeur du traitement des données (de plus complexes et lourdes dans le cas de CLAPI, exigeant un travail de plus en plus important et de plus en plus de technique) et son coût font ainsi naître le risque paradoxal d'un rétrécissement du champ des recherches sur des corpus « rentables », parce que très standard et ne posant pas de problèmes de droit. S'il a été à un certain moment essentiel de construire les conditions du partage des données, il convient maintenant de protéger la possibilité de lignes de recherche qui n'alimentent pas directement les infrastructures ainsi mises en place, mais qui contribuent néanmoins tout autant à l'enrichissement du panorama de la recherche sur l'oral et sur la langue parlée en interaction.

BIBLIOGRAPHIE

Baldauf-Quilliatre H. (2014a). « Répétition et encouragement », *Semen* 38 [Véronique Magri-Mourgues / Alain Rabatel (éd.) : Pragmatique de la répétition], 115-135.

Baldauf-Quilliatre H. (2014b). « Formate knapper Bewertungen beim empraktischen Sprechen », *in* C. Schwarze, C. Konzett (éd.) *Hinter den Kulissen : Aktuelle Projekte aus der Interaktionsforschung – methodologisch betrachtet*. Frankfurt : Lang, 107-130.

Balthasar L. & Bert M. (2005). « La plateforme "Corpus de langues parlées en interaction" (CLAPI) », *Lidil* 31 : 13-33.

Baude O. (éd.) (2006). *Corpus oraux, guide des bonnes pratiques 2006*. Paris & Orléans : Éditions du CNRS & Presses universitaires d'Orléans.
https://hal.archives-ouvertes.fr/hal-00357706/file/Corpus_Oraux_guide_des_bonnes_pratiques_2006.pdf

Bruxelles S. & Traverso V. (2003). « Les corpus de langue parlée en interaction au GRIC », *in* D. Pusch & F. Raible (éd.) *Romanistiche Korpuslinguistik*. Tübingen : Gunter Narr Verlag, 59-70.

Colón de Carvajal I., Lascar J. & Traverso V. (à paraître). « Et l'impact de la caméra alors... », Revue en ligne Ethnographiques.org.

Colón De Carvajal I. (2013). « Du corpus enregistré au corpus analysé : questions méthodologiques sur l'utilisation d'outils de requêtes informatisés. Corpus, Données, Modèles », *Cahiers de Praxématique* 54-55/2010, Montpellier : PULM, 313-326. [halshs-00630514].

Étienne C. (2009). « La TEI dans le Projet CLAPI, Corpus de langues parlées en interaction », *TEI Council*, Lyon.

Groupe ICOR (L. Balthasar, S. Bruxelles, L. Mondada, V. Traverso) (2007). « Variations interactionnelles et changement catégoriel : l'exemple de 'attends' », *in* Auzanneau M. (éd.) *La Mise en œuvre des langues dans l'interaction*. Paris : L'Harmattan, 299-319.

Groupe ICOR (M. Bert, S. Bruxelles, C. Étienne, L. Mondada, S. Teston-Bonnard, V. Traverso) (2008a). « 'Oh::, oh là là, oh ben...', les usages du marqueur 'oh' en français parlé en interaction », *in* J. Durand, B. Habert & B. Laks (éd.) *Congrès mondial de linguistique française*. Paris, France. En ligne, <10.1051/cmlf08099>. <halshs-00356377>

Groupe ICOR (M. Bert, S. Bruxelles, C. Étienne, L. Mondada, V. Traverso) (2008b). « Tool-assisted analysis of interactional corpora : voilà in the CLAPI database », *Journal of French Language Studies* 18 (1) : 121-145.

Groupe ICOR (M. Bert, S. Bruxelles, C. Étienne, 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* 33 (2) : 243-268.

Groupe ICOR (M. Bert, S. Bruxelles, C. Étienne, L. Mondada, V. Traverso) (2010). « Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL) », *Pratiques* 147-148 : 17-34.

Groupe ICOR (C. Étienne, S. Bruxelles, E. Jouin, L. Mondada, F. Oloff, V. Traverso) (à paraître). « Phénomènes et unités : questions autour de la détection automatique des répétitions dans un corpus de langue parlée en interaction », *in (DES-)Organisation de l'oral de la segmentation à l'interprétation*. Rennes.

Groupe ICOR (H. Baldauf-Quilliatre, S. Bruxelles, S. Diao-Klaeger, E. Jouin-Chardon, V. Traverso) (2014). « Oh là là : the contribution of the multimodal database CLAPI to the analysis of spoken French », *in* H. Tyne, V. André, A. Boulton, C. Benzitoun, Y. Greub (éd.) *Ecological and Data-Driven Perspectives in French Language Studies*. Newcastle : Cambridge Scholars Publishing, 167-198.

Gülich E. (2008). « Le recours au préformé : une ressource dans l'interaction conversationnelle », *in* J. Durand, B. Habert & B. Laks (éd.) *Congrès mondial de linguistique française*. Paris, France. Disponible en ligne sous : http://www.linguistiquefrancaise.org/index.php?option=article&access=doi&doi=10.1051/cmlf08315.

Laurier E. & Philo C. (2006). « Natural problems of naturalistic video data », *in* H. Knoblauch, J. Raab, H.-G. Soeffner & B. Schnettler (éd.) *Video-Analysis Methodology and Methods, Qualitative Audiovisual Data Analysis in Sociology*. Oxford : Peter Lang, 183-192.

Mondada L. (2006). « Video Recording as the Reflexive Preservation and Configuration of Phenomenal Features for Analysis », *in* H. Knoblauch, J. Raab, H.-G. Soeffner & B. Schnettler (éd.) *Video-Analysis Methodology and Methods, Qualitative Audiovisual Data Analysis in Sociology*. Oxford : Peter Lang, 51-68.

Mondada L. (2007). « Multimodal ressources for turn-taking : Pointing and the emergence of possible next speakers », *Discourse Studies* 9/2 : 195-226.

Mondada L. (2012), « Talking and driving : Multiactivity in the car », *Semiotica* 191, 223-256.

Potter J. (2006). « Naturalistic Data », *in* V. Jupp (éd.) *The Sage Dictionary of Social Research Methods*. London : Sage. Brockington.

Schmale G. (2013). « Formen und Funktionen vorgeformter Konstruktionseinheiten in authentischen Konversationen / Forms and Functions of Formulaic Construction Units in Conversation », *Linguistik Online* 62, 5/2013, http://www. linguistik-online.de/62_13/.

Ticca A. C. & Traverso V. (à paraître, 2015). « Territoires corporels, ressenti et paroles d'action : des moments délicats de la consultation médicale avec interprète », *Langage et Société*.

Traverso V. (2011). « Analyser un corpus de langue parlée en interaction : questions méthodologiques », *Verbum* 4 : 313-329.

Traverso V. (2012a). « 'Le salon bibliothèque' : délimitation et partage des espaces. Usage des annonces dénominatives désignatives dans la visite guidée », *in* J.-P. Dufiet (éd.) *Les Visites guidées. Discours, interaction, multimodalité*. Trento : Presses de l'Université de Trento, 55-85.

Traverso V. (2012b). « Analyses interactionnelles : repères, questions saillantes et évolution », *Langue Française* 175 : 3-17.

Traverso V. *et al.* (2012). « Analyses de l'interaction et linguistique : état actuel des recherches en français », *Langue française* 175.

Traverso V. (2014). « La construction de (l'attention visuelle sur) l'objet au cours de la visite guidée : étude d'un cas limite », *in* J. P. Dufiet (éd.) *L'Objet d'art et de culture à la lumière de ses médiations*. Trento : Coll. Labirinti, 43-85.

ANNEXES

Occurrences de « trop »

Le total des occurrences de « trop » dans ces 24 corpus s'élève à 337 occurrences (la base de données en contient au total 547).

Corpus	Année	durée min	nombre d'occ.	Trop Excès	Trop intensif
Négociation sur les loyers – commission de conciliation,	1984	115	18	9	9
Mode – interactions sur un thème imposé,	1982	22	5	4	1
Français des années 80 – entretiens sociolinguistiques,	1984	72	32	27	5
Conversations familières – Visites	1985-1990	61	13	11	2
Interactions commerciales – bureau de tabac presse	1986	120	1	1	
Enquête de sociologie urbaine –Paris Marais,	1989-1990	171	19	15	4
Conversations téléphoniques en entreprise	1997	25	4	3	1
Débat sur l'immigration – TP d'étudiants	1997	78	8	8	0
Négociation sur le partage de biens – notaires	1997-1998	36	2	1	1
Interactions commerciales – vente à domicile encyclopédies,	1998-1999	6	1	0	1
Interactions dans un commerce – magasin de retouches (papotages)	2001	22	6	4	2
Réunion de conception en architecture – Mosaic	2002	78	20	12	8
Consultations chez les dentistes	2003	35	3	1	2
Réunion de travail entre publicitaires – Lyon Saxe	2004	58	7	5	2
Repas. Conversations entre étudiants	2006	47	36	14	22
Interactions pendant la tournée de facteurs	2006-2007	24	3	3	0
Repas Épinards	2008	31	8	2	6
Repas Kiwi	2008		88	2	86
Repas Olives	2008	29	20	2	18
Conversations en ligne	2007-2008	14	5	4	1

Session de jeux vidéo entre jeunes	2007	106	28	16	12

NOTES

1. http://icar.univ-lyon2.fr/projets/corinte/

2. Dans cette perspective, les entretiens sont des données provoquées et ils illustrent un genre interactionnel spécifique. Le choix a été fait que la base de CLAPI ne contienne pas de données médiatiques pour des questions de droit (des données radiophoniques ont en revanche été collectées dans le projet CIEL-F, et elles sont hébergées dans CLAPI-CIELF, voir 2.3).

3. http://icar.univ-lyon2.fr/projets/corinte/bandeau_droit/convention_icor.htm

4. http://icar.univ-lyon2.fr/projets/corinte/

5. http://icar.univ-lyon2.fr/projets/corvis/

6. http://talkbank.org/

7. Étienne, 2009.

8. http://www.ciel-f.org/

9. http://ircom.huma-num.fr

10. https://www.ortolang.fr

11. http://www.projet-orfeo.fr

12. Cet usage est considéré comme familier dans le TLFI, avec l'exemple « Il en avait de trop à bouffer le général, puisqu'il touchait d'après le règlement quarante rations pour lui tout seul » (Céline, *Voyage*, 1932, p. 33).

13. http://clapi-fpie.ish-lyon.cnrs.fr

RÉSUMÉS

Dans cette contribution, nous présentons la base CLAPI développée au laboratoire ICAR dans le contexte de l'évolution des bases de données de langues parlées en France au cours des trente dernières années. Nous détaillons les deux composantes de CLAPI, l'archive de corpus de langue parlée en interaction audio et vidéo enregistrés dans des situations sociales naturelles variées, et la plateforme d'outils.

L'usage et l'apport de CLAPI sont illustrés par deux études. L'une décrit comment la base peut être utilisée pour des travaux de linguistique interactionnelle intégrant la multimodalité (« oh là là ») ; l'autre concerne une recherche combinant données et métadonnées (« trop »).

L'article est aussi l'occasion d'un bilan plus général. La mise en perspective montre en effet qu'après la période des questions est venue celle des dilemmes. La période des questions, choix et décisions à toutes sortes de niveaux a accompagné la mise en place des bases de données. L'expérience permet maintenant de mesurer leurs indéniables apports en termes non seulement de quantité de données disponibles (et traitables grâce aux outils), mais aussi de qualité (comme conséquence des exigences de standardisation liées au partage des données). La période des dilemmes nous conduit à nous interroger sur les meilleurs choix à opérer aujourd'hui dans les relations entre la poursuite des recherches sur des corpus variés (et parfois sensibles) et les exigences des bases de données ouvertes.

In this contribution, we present the development of the CLAPI by the ICAR Lab in the context of the evolution of the databases of spoken languages in France during the last thirty years. We describe the two components of CLAPI, the archive of corpus of spoken languages in interaction, audio and video, recorded in varied naturally-occurring social situations, and the plateform of tools.
The use and the support of CLAPI the research are shown out of two studies. One illustrates how the database can be used for working in an interactional linguistic perspective, including multimodality (“oh là là”); the other concerns a research combining data and metadata (“trop”).
The article is also the occasion of a more general assessment. The perspective on the last thirty years shows that after a period of questions came that of dilemmas. The period of questions, choices and decisions at various levels accompanied the implementation of the databases. The experience enables now to measure their undeniable contributions in terms not only of quantity of available data (and possibly dealt with supported by the tools), but also of quality of the data (as a consequence of the requirements of standardization linked to the needs of sharing the data). The period of the dilemmas leads us to wonder about the best choices to be operated today among continuing research on varied corpuses (sometimes delicate) and the requirements of the databases.

INDEX

Mots-clés : parole en interaction, multimodalité, interopérabilité, banques de données
Keywords : talk-in-interaction, multimodality, interoperability, databank

AUTEURS

H. BALDAUF-QUILLIATRE

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2

I. COLÓN DE CARVAJAL

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2

C. ETIENNE

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2

E. JOUIN-CHARDON

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2

S. TESTON-BONNARD

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2

V. TRAVERSO

Groupe ICOR, UMR 5191 – CNRS / Université Lyon 2