Pour une archive des langues parlées en interaction
Statuts juridiques, formats et standards, représentativité

Responsable	Christian Plantin
Co-reponsable	Lorenza Mondada
Laboratoire de rattachement	ICAR (UMR5191 CNRS, Univ. Lyon2, ENS LSH)
Membres d'ICAR engagés dans le projet	Lukas Balthasar, Michel Bert, Sylvie Bruxelles, Lorenza Mondada Christian Plantin, Véronique Traverso
Thématique de recherche	Linguistique interactionnelle, droit, informatique
Partenaires et leurs Laboratoires de rattachement	LACITO, Paris (Michel Jacobson) LORIA, Nancy (Matthieu Quignard, Laurent Romary) GREYC, Caen (Anne Nicolle) LIMSI, Paris (Laurence Devillers) LIUM, Le Mans (Daniel Luzzati) RIM, Saint-Etienne (Jean-Jacques Girardot)

1. La situation des archives de langues parlées en interaction en France

La constitution d'archives de langue parlée en interaction est une exigence ressentie de manière générale sur la scène internationale depuis plusieurs décennies. C'est ainsi que dans plusieurs pays ont été constituées des banques de données de corpus de données interactionnelles : tel est le cas par exemple de la partie orale du BNC (British National Corpus, http://www.natcorp.ox.ac.uk/), du corpus de langue parlée de l'Institut für deutsche Sprache (http://www.ids-mannheim.de/ksgd/), ou encore du corpus national de néerlandais parlé (http://lands.let.kun.nl/cgn/ehome.htm). En francophonie, il n'existe actuellement aucune initiative d'envergure comparable, mais des initiatives locales généralement liées à des projets spécifiques (en France, PFC http://infolang.u-paris10.fr/pfc, DELIC www.up.univ-mrs.fr/delic/ ; en Belgique VALIBEL http://valibel.fltr.ucl.ac.be/ ou ELICOP http://bach.arts.kuleuven.ac.be/elicop/). La banque de corpus CLAPI développée au Laboratoire ICAR (http://icar.univ-lyon2.fr/) présente la particularité de recueilllir des enregistrements de parole en interaction recueillis dans leur contexte social ordinaire de production et d'être ouverte à la fois à des corpus issus des recherches du laboratoire et de corpus pouvant être déposés par tout chercheur soucieux d'archiver et de mettre à disposition ses données à des conditions particulières (cf. Balthasar & Bert 2005 pour une présentation).

Ce type d'archive - ou banque de données de corpus - poursuit différentes visées :

- une visée *patrimoniale* : sauvegarder les « façons de parler » que l'on peut documenter au fil du temps, dans toute culture, majoritaire ou minoritaire, en constituant ainsi progressivement une documentation historique des usages de la langue en interaction,
- une visée *scientifique* : ces corpus constituent la base empirique à partir de laquelle effectuer des études de phénomènes linguistiques divers - description de la grammaire, du changement linguistique, de l'organisation interactionnelle - mais touchant aussi à d'autres problématiques disciplinaires - la culture communicationnelle, la dynamique des relations sociales, l'histoire des pratiques ordinaires, institutionnelles, professionnelles, etc.
- une visée *appliquée* : ces corpus sont une ressource importante pour l'enseignement des langues et pour la formation dans des domaines aussi variés que les activités en équipe ou les pratiques de management de la discussion ; ainsi que pour la construction de dictionnaires, voire des applications industrielles telles que la simulation ou l'automatisation de dialogues.

Historiquement, les corpus de langue parlée en interaction ont été nettement moins privilégiés que les corpus écrits : cela s'explique par le fait que les corpus de textes sont plus faciles à constituer, grâce à l'énorme disponibilité de textes dans les bibliothèques et sur Internet ; et que les corpus de l'oral posent des problèmes spécifiques qui en complexifient le recueil. Ceux-ci en effet nécessitent - lorsque l'objectif est de documenter des activités langagières dans des contextes sociaux authentiques variés - des enquêtes de terrain et des dispositifs d'enregistrements complexes, et produisent des documents textuels - les transcriptions - qui n'obéissent pas aux mêmes standards que les textes écrits et qui ne sont donc pas exploitables par les mêmes moteurs de recherche et outils (semi)automatiques d'analyse.
Les corpus de langue parlée en interaction sont en effet des objets complexes, qui comprennent :
- des *données primaires*, sous forme d'enregistrements audio ou vidéo d'activités langagières, auxquels peuvent s'ajouter des documents, textes, images, objets, traces informatiques mobilisés ou produits durant ces mêmes activités.
- des *données secondaires*, sous forme de transcriptions (éventuellement sous la forme de différentes versions), auxquelles s'ajoutent les conventions de transcription, les autorisations signées des participants, d'éventuelles notes ethnographiques, des annotations du corpus, ainsi que les descripteurs de l'enregistrement (les métadonnées).
La complexité des corpus relève de cette articulation entre des objets matériellement et épistémologiquement différents, pouvant proliférer (on peut envisager autant de transcriptions que d'objets de recherche et d'appartenances théoriques) et nécessitant néanmoins une gestion unifiée et donc standardisée. De cette complexité naissent plusieurs questions qui ont constitué le socle du projet « *Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité ».*

2. Les problématiques du projet : qu'est-ce que la « qualité » d'un corpus ?
Bien que la volonté de constituer des corpus de langue parlée en interaction se fasse sentir de plus en plus largement, les motivations scientifiques qui la soutendent sont souvent hétérogènes, produisant des exigences diverses en matière de définition des corpus visés.
Ainsi, la banque de corpus CLAPI émane d'une communauté scientifique particulière, celle de la linguistique interactionnelle, qui pour accomplir son programme d'analyse des pratiques situées des locuteurs a procédé très tôt à la constitution de corpus de « naturally occurring interactions » (cf. ten Have, 1998). L'exigence de « naturalité » - c'est-à-dire de pouvoir observer des activités ancrées dans leur lieu social ordinaire d'effectuation, telles qu'elles se déroulent habituellement que le chercheur soit présent ou non - émane de la reconnaissance du caractère *localement situé* de l'interaction et de son organisation accomplie moment par moment par les participants en tenant compte des *contingences* de l'interaction et en interprétant les détails des actions des uns et des autres afin de s'y ajuster mutuellement. De cette vision de la parole-en-interaction découle donc l'exigence de recueillir les données interactionnelles en contexte.
D'autres courants théoriques sont porteurs d'autres exigences : par exemple ceux qui attribuent moins d'importance au contexte - en considérant que ses variations et contingences n'interviennent que de manière marginale dans l'organisation de la parole - et qui réclament des données d'une haute qualité sonore privilégieront des interactions étroitement contrôlées par le chercheur et se déroulant dans des conditions optimales en laboratoire.
Il est néanmoins intéressant de noter que de plus en plus de courants théoriques préfèrent, s'ils en ont le choix, de travailler sur des données empiriques naturelles plutôt que sur des données fabriquées par introspection ou par simulation ou expérimentation. Cela est le cas aussi des approches de la langue qui *ne* souscrivent *pas* à un modèle interactionniste. Dès lors, les

données produites par ce dernier semblent fournir un standard d'exigences élevées, censées intéresser d'autres approches aussi.
La question se pose donc de savoir s'il est possible d'énoncer des critères permettant d'évaluer la « qualité » d'un corpus - en considérant que n'importe quel enregistrement de paroles orales en interaction ne constitue pas en soi et d'emblée un corpus intéressant à archiver. Le projet « *Pour une archive des langues parlées en interaction* » s'est penché sur cette question en s'appuyant sur les expériences effectuées dans le cadre de la constitution de la base CLAPI et en isolant une série de dimensions certes imbriquées mais relativement autonomes :
- *la question juridique* : pour pouvoir exploiter les corpus et *a fortiori* les partager et les diffuser, il faut que les enregistrements aient été effectués dans des conditions juridiquement adéquates, avec les accords de toutes les parties concernées. La qualité d'un corpus dans ce sens est non seulement juridique - ce qui en garantit l'exploitabilité légale - mais aussi éthique, posant la question des modalités d'établissement et d'entretien des relations entre les chercheurs et les sujets de leurs observations.
- *la question technique* : la qualité technique des corpus n'est pas seulement importante pour en garantir la conservation, mais déjà pour en permettre l'exploitation : ainsi par exemple un corpus numérisé avec un taux de compression élevé risque de gommer des détails qui seraient utiles à certaines analyses ; de même, un corpus numérisé dans un format non standard risque de n'être lisible que par celui qui l'a numérisé.
- *la question sociolinguistique* : la qualité sociolinguistique du corpus tient aux exigences qui ont présidé à son enregistrement, à la fois dans l'approche de terrain qui l'a rendu possible et dans le dispositif d'enregistrement lui-même. Dans ce sens, la qualité sociolinguistique d'un corpus ne relève pas uniquement de son « authenticité », mais aussi de la qualité de la relation aux participants et de l'adéquation du dispositif technique d'enregistrement à la situation.
Nous allons expliciter ci-dessous quelques résultats auxquels a abouti le projet sur ces trois points.

3. Enjeux juridiques
Progressivement, et de manières différenciées selon les disciplines, les objets d'enquête et les traditions nationales, les enjeux juridiques se sont imposés comme une des dimensions constitutives des processus de constitution de corpus de langue parlée en interaction.
En outre, cette dimension juridique concerne toutes les étapes qui marquent le processus d'enquête et de recherche, devenant ainsi un très bon indicateur des procédures adoptées au fil du travail scientifique dans son ensemble.
Les problèmes juridiques concernent notamment trois sphères :
- le *respect de la vie privée* des enquêtés : cet aspect est central dès qu'il s'agit d'enregistrer en audio ou vidéo les pratiques des personnes dans leur vie ordinaire.
- le *droit d'auteur* : cet aspect invite à problématiser la notion d' « auteur » pour situer les acteurs intervenant dans la constitution du corpus (collecteurs, transcripteurs, etc.).
- les droits et obligations liés à la *gestion des banques de données* : cet aspect concerne les conventions à établir entre les responsables de corpus, la banque de données et les personnes désirant la consulter.
Nous synthétiserons quelques-uns de ces points en distinguant différentes étapes parcourues lors de la constitution et exploitation des corpus :
- le *recueil des données* : lors de la phase préliminaire de l'arpentage du terrain, de la prise de contact des participants ou des informateurs, du choix des activités et des moments à enregistrer, l'enjeu juridique et éthique s'incarne dans l'organisation pratique des modalités d'approche des participants - qui peuvent dès lors être abordés en tant qu'« objets », « sujets », « informateurs », « locuteurs », « acteurs sociaux », « participants », « partenaires

du projet de recherche » avec des effets structurants forts différents sur la relation « enquêteur » / « enquêté ». Avant l'enregistrement et en prévision de celui-ci, cela se traduit dans la construction du *consentement éclairé* des enquêtés : ce terme articule l'autorisation donnée par les participants à l'enquête avec leur information et compréhension des visées du projet dont émane l'enregistrement. Une fois obtenu l'accord des participants - qui peut être oral ou écrit, et qui dans des sociétés lettrées est souvent signé - l'enregistrement matérialise lui-même certaines des contraintes énoncées dans les documents d'autorisation (par ex. les limitations concernant les moments enregistrables ou l'interdiction de tourner en vidéo). Le dispositif d'enregistrement lui-même - son caractère caché ou visible, intrusif ou discret, avec un cadrage total de la scène ou le ménagement d'angles morts qui permettent un retrait pour les participants - intègre des éléments qui relèvent de l'accord juridique obtenu comme du respect éthique de la vie privée des participants.
- la numérisation et préparation des données primaires : lors de cette phase, il s'agit notamment de procéder à *l'anonymisation* des bandes enregistrées (« beepage » de l'audio et « floutage » de la vidéo).
- l'élaboration des données secondaires : a) transcription, b) métadonnées, c) annotations. Lors de ces différentes phases, la question du respect de la vie privée concerne l'anonymisation des transcriptions, l'évitement de la possibilité de recouper des informations de sources différentes (entre le corps des transcriptions et les métadonnées notamment), l'image des enquêtés que produit le corpus ainsi élaboré (évitement de stéréotypes, discrimination, diffamation des enquêtés). Lors de cette phase se multiplient aussi les personnes travaillant sur le corpus et pouvant prétendre au droit d'auteur : cela concerne les concepteurs du dispositif d'enregistrement s'il est original, du dispositif de montage, de la transcription, de l'annotation, de l'exploitation analytique des données.
- la diffusion des données au-delà de l'équipe et du projet initiaux (avec un éventuel changement de finalités, impliquant une nouvelle autorisation de la part des participants) : lors de cette phase se posent notamment des problèmes d'explicitation des droits et des obligations concernant les responsables du corpus, les responsables de la base de donnes et les utilisateurs.
Ces réflexions ont donné lieu, durant le projet, à une participation importante du laboratoire ICAR à la rédaction d'un *Guide des bonnes pratiques juridiques en linguistique*, en collaboration avec la DGLF (à paraître ; voir la version provisoire sur le site web http://www.culture.gouv.fr/culture/dglf/cotelecharger.htm).

4. Enjeux techniques

La qualité du corpus est souvent entendue au sens technique du terme, au sens de sa qualité *sonore* - qui permet non seulement d'effectuer des transcriptions fines mais aussi de soumettre l'enregistrement à des analyses informatiques prosodiques et acoustiques - et de sa qualité *visuelle* - qui permet, grâce à des modes de compression adéquats, de suivre précisément, image par image, la trajectoire d'un geste, de noter un regard, de déchiffrer ce que vient d'écrire un participant, etc.
Mais ces questions de qualité - concernant les formats audio et vidéo, les codecs utilisés pour la compression, les outils servant à l'alignement des enregistrements et des transcriptions, les standards d'annotation - dépendent de manière plus radicale du début de la chaîne de production d'un corpus, c'est-à-dire de la conception du dispositif d'enregistrement lui-même, voire de l'approche du terrain en vue de l'enregistrement. Celles-ci présupposent immédiatement une analyse ethnographique, sociologique et linguistique du contexte où a lieu l'enregistrement, qui permet d'effectuer les choix techniques. Ainsi en est-il par exemple du placement des micros et des caméras, qui posent avant tout des problèmes d'analyse de l'activité, obligeant de prévoir les mouvements que feront les participants et de cadrer ainsi

adéquatement l'image, ainsi que de prévoir les problèmes sonores (localisation de sources de musique ou de bruits).
Durant le projet, le laboratoire ICAR a réalisé des expériences en matière d'enregistrement que nous avons appelé « multiscope » (Balthasar & Mondada, à paraître) : il s'agit de disposer plusieurs caméras, de manière à couvrir le même espace de différentes perspectives ou bien de couvrir des espaces complémentaires entre lesquels les participants peuvent se déplacer. Ces caméras permettent ainsi un suivi continu des détails de l'action - que ce soit des regards, des mouvements, des gestes. Dans le dispositif retenu dans ces expériences, le choix a aussi été fait d'alimenter les caméras sur secteur et d'enregistrer directement l'image sur disque dur, évitant ainsi les batteries et les cassettes qui limitent l'autonomie des équipements : cela a permis de réaliser des enregistrements de 6-7 heures d'affilée sans interruption et en absence de l'enquêteur - de manière minimalement invasive pour les participants et non perturbante pour le déroulement de l'activité (qui est autrement suspendu et transformé par le changement des cassettes). Plusieurs corpus ont été réalisés de cette manière : l'un à Bruxelles en collaboration avec des historiens interrogeant des témoins ayant vécu dans les camps de concentration durant la 2e guerre mondiale, nous ayant permis de filmer un entretien effectué en studio (6h) tout en imposant une discrétion absolue de notre part en ce qui concerne les équipements utilisés. Les autres corpus réalisés sont des repas entre amis enregistrés au domicile des hôtes, dans différents contextes sociaux et culturels (des professionnel, des familles, des étudiants). Ces corpus ont notamment permis de développer une réflexion sur le type d'observabilité et d'analyse que les données rendent possibles (Mondada, 2005 ; Mondada & Traverso 2005).
Ces considérations montrent l'articulation indissociable des problèmes techniques, des approches du terrain et des enjeux analytiques, qui se trouvent imbriqués dans les problèmes sociolinguistiques que nous aborderons ci-dessous.

5. Enjeux sociolinguistiques

La qualité d'un corpus se décline non seulement du point de vue de la sauvegarde des droits des enquêtés et de la qualité technique de la prise de son et de vue, mais aussi du point de vue des dimensions variablement catégorisées dans la littérature par des termes comme « authenticité », « spontanéité », « naturalité », « représentativité ». Dans la littérature émanant de l'analyse conversationnelle, le terme qui a été privilégié est celui de « naturalité » qui renvoie à un « naturalisme » de l'observation, portant sur des situations qui ne sont pas construites, orchestrées, organisées par le chercheur pour les fins de son enquête. Au contraire, celle-ci s'ajuste aux contraintes du contexte et des actions des participants, en s'efforçant de ne pas les transformer par ses propres exigences techniques ou scientifiques. Dans ce sens, le terme de « naturalité » peut être rapproché de celui d' « authenticité ». Le terme de « spontanéité » est par contre évité, car il introduit une évaluation sur la base de la valorisation d'un certain état psychologique et il exclut d'emblée les situations où les participants sont contraints socialement ou institutionnellement par un certain degré de formalité - dans des situations tout à fait « naturelles ».
La question de la « représentativité » se pose différemment : elle repose sur une démarche consistant à effectuer un échantillonnage selon des catégories sociolinguistiques prévues a priori par l'enquête ou la banque de données (Biber, 1995). La démarche de l'analyse conversationnelle se distingue pour différentes raisons, notamment par le fait qu'elle privilégie des catégories « émiques » plutôt qu' « étiques », c'est-à-dire des catégories rendues pertinentes par les participants eux-mêmes (Sacks, 1972) plutôt que par les chercheurs en vertu de leurs typologies ou leurs modèles. Le caractère « représentatif » dans cette optique est produit par les participants eux-mêmes, qui s'orienteront vers une scène, une action, ou un événement traité comme relevant « typiquement » d'une certaine catégorie.

Ce privilège des données naturelles n'exclut pas la prise en compte *réflexive* (Rabinow, 1977 ; Woolgar, 1988) des effets configurants du micro ou de la caméra : dans cette perspective, il s'agit moins de traiter l'équipement technique comme un « biais » (cf. le « paradoxe de l'observateur » cher à Labov et les procédés qu'il met en oeuvre pour l'éviter, 1984) néfaste pour l'enquête, mais comme un élément - souvent traité comme un véritable participant - qui fait partie de la situation et que les participants peuvent non seulement rendre pertinent mais exploiter comme ressource pour l'organisation de leurs affaires. Ces considérations invitent à traiter l'orientation vers la caméra comme une dimension constitutive, agisssant comme un révélateur des modes d'organisation des activités dans ce contexte particulier - rendant visible non seulement le caractère structuré et structurant de l'enquête mais aussi les procédés de structuration de l'activité en cours par les participants.

6. Conclusion

Au terme de ce bref parcours, la notion de « qualité » d'un corpus apparaît dans sa complexité comme articulant des enjeux juridiques, techniques et sociolinguistiques qui apparaissent souvent comme étant étroitement imbriqués les uns aux autres. Leur définition, reconnaissance, contrôle et théorisation par le chercheur dépend de la mentalité analytique qui motive le recueil des données et, en définitive, le traitement auxquelles elles seront soumises.

7. Références citées

Balthasar, L. & Bert, M. (2005). La base de données « Corpus de langues parlées en interaction » (CLAPI) : genèse, état des lieux et perspectives. In Savelli, M. (éd.). Corpus oraux et diversité des approches. Numéro spécial de *Lidil*, 31.

Balthasar, L. & Mondada, L. (à paraître). « Multiscope videos », *Proceedings of the 2d ISGS Conference, Interacting Bodies, Lyon, june 15-18, 2005*.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge Universty Press.

Have, P. ten (1998). *Doing Conversation Analysis. A Practical Guide*. London: Sage.

Labov, W. (1984). Field Methods of the Project on Linguistic Change and Variation. In J. Baugh & J. Sherzer (Eds.), *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs, NJ: Prentice-Hall.

Mondada, L. (2005). L'analyse de corpus dans la perspective de la linguistique interactionnelle : des analyses de cas singuliers aux analyses de collections. In.:A. Condamine (éd.), *Sémantique et corpus*, Paris : Hermès, 76-108.

Mondada, L. & Traverso, V. (2005). (Dés)alignements en clôture : une étude interactionnelle de corpus de français parlé en interaction. In Savelli, M. (éd.). Corpus oraux et diversité des approches. Numéro spécial de *Lidil*, 31.

Rabinow, P. (1977). *Reflexions on Fieldwork in Morocco*. Berkeley: University of California Press.

Sacks, H. (1972). An initial investigation of the usability of conversational materials for doing sociology. In D. Sudnow (Ed.), *Studies in Social Interaction* (pp. 31-74). New York: Free Press.

Woolgar, S. (Ed.). (1988). *Knowledge and Reflexivity: New frontiers in the Sociology of Knowledge*. London: Sage.