CORPUS ORAUX : LES *BONNES PRATIQUES* D'UNE COMMUNAUTÉ SCIENTIFIQUE

Olivier BAUDE
CORAL, Université d'Orléans
Délégation Générale à la Langue Française et aux Langues de France

SOMMAIRE
0. Introduction
1. Contextes pour une diffusion de la recherche
1.1. La linguistique de corpus et l'oral
1.2. Une politique de diffusion
1.3. Les initiatives de mutualisation
1.4. Le guide des bonnes pratiques
2. Aspects juridiques
2.1. Définition de l'objet
2.2. Domaines juridiques concernés
2.3. Diffusion scientifique et droit d'auteur
3. Éléments de réponses
3.1. Expliciter la démarche du chercheur
3.2. Le recueil de consentement
3.3. L'anonymisation
3.4. Structure du corpus
4. Conclusion

0. Introduction

Les problèmes juridiques liés à la diffusion des corpus oraux ont été l'occasion d'une démarche originale adoptée par une communauté scientifique ouverte à un travail pluridisciplinaire. Cette démarche a comporté plusieurs étapes. Une lecture croisée des textes juridiques par les linguistes et les juristes a permis de repérer les problèmes. Les chercheurs ont ensuite accepté d'expliciter leurs pratiques au regard de la législation. Cette étape fondée sur la réflexivité a permis d'élaborer des propositions pour de bonnes pratiques partagées par la communauté scientifique et de repérer des aspects juridiques qui posent des difficultés dans l'état actuel du droit.
Ce travail s'est concrétisé par la rédaction de l'ouvrage *Corpus oraux, guide des bonnes pratiques 2006*[1]. Rédigé par un groupe de travail constitué de linguistes, juristes, informaticiens et conservateurs, cet ouvrage a pour vocation explicite, d'éclairer la démarche des chercheurs, de repérer les problèmes et les solutions juridiques et de favoriser l'émergence de pratiques communes pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux.
Le résultat de ce travail interdisciplinaire ouvre les portes d'une réflexion sur les pratiques des chercheurs en sciences sociales et leurs relations aux données, à l'heure de l'exploitation et de la diffusion en masse de celles-ci.

1. Contextes pour une diffusion de la recherche

1.1. La linguistique de corpus et l'oral

Depuis plus de 30 ans le domaine de la linguistique de corpus s'est considérablement développé autour des corpus écrits, aussi bien en ce qui concerne la masse des données disponibles que l'élaboration d'outils de traitement automatique de celles-ci. La situation est totalement différente pour les corpus oraux[2]. Pourtant, les toutes nouvelles technologies en matière de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils (transcriptions synchronisées sur le signal, annotations, etc.) ouvrent des perspectives prometteuses pour les études sur les corpus de langues parlées. De nombreux corpus ont été constitués ou sont en

[1] *Corpus oraux, guide des bonnes pratiques 2006*, Paris, CNRS éditions.
[2] Pour plus de commodités et selon l'usage, nous utiliserons les termes *corpus oraux* comme termes génériques définissant des collections ordonnées d'enregistrements de productions linguistiques orales et multimodales.

cours de constitution et leur diffusion pose des problèmes juridiques et éthiques que la communauté scientifique doit prendre en charge. Pourquoi et comment ?

1.2. Une politique de diffusion

Depuis 1982 et la loi pour la recherche et le développement technologique en France[1], la diffusion des résultats fait partie des missions des chercheurs. Plus récemment, la déclaration de Berlin signée par la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) le 22 octobre 2003 plaide pour la constitution de bases de connaissances en libre accès[2]. Enfin, les programmes de numérisation patrimoniale comprennent un volet de valorisation des ressources numérisées (cf. texte de Lund de 2001 prônant la mise en place des standards d'interopérabilité).

1.3. Les initiatives de mutualisation

Cette dernière notion de standards d'interopérabilité se retrouve dans différentes initiatives internationales (TEI, groupe de travail ISO TC37 SC4 pour la gestion des ressources linguistiques, protocole d'échange OAI, norme ANSI/NISO Z39.50, projet Open Language Archive Community, etc.) ainsi que dans des choix techniques (utilisation du langage de balisage XML par exemple). Dans le même cadre de valorisation de la recherche et de mutualisation des ressources, le CNRS s'est doté, en 2005, d'une direction de l'information scientifique, et développait un an plus tard des centres de ressources numériques.
Dans le même temps des laboratoires de recherche lançaient différentes initiatives pour la diffusion et l'accessibilité des corpus oraux (Base Clapi du laboratoire Icar[3], projet Corpus Oraux de l'EPML 50[4], programme Archivage du Lacito[5], constitution de grands corpus disponibles comme le projet Phonologie du Français contemporain[6], C-oral-Rom[7], etc.).

1.4. Le Guide des bonnes pratiques

C'est dans ce contexte que la Délégation générale à la langue française (direction du ministère de la culture) et le CNRS ont constitué un groupe de travail pluridisciplinaire qui a pour mission de favoriser la collecte et l'exploitation de corpus oraux.
Ce groupe de travail comporte des linguistes experts et des chercheurs de "terrain" porteurs de projets actuels, des représentants des fédérations de laboratoire du CNRS, des juristes, des représentants des grands organismes de conservation sous la tutelle du Ministère de la Culture et des juristes de ces institutions. L'objectif premier était de permettre un travail en commun sur un objet scientifique, de favoriser sa conservation et surtout sa diffusion (diffusion auprès de différentes équipes de recherche mais aussi auprès d'un public plus large). Or, il est très vite apparu que les aspects juridiques étaient les premiers obstacles à la diffusion de l'oral transcrit (qui est propriétaire de quoi ? Qui est responsable de la diffusion ? Quelles sont les autorisations à recueillir ? Qu'en est-il du droit d'auteur ?, etc.). Enfin, ce travail sur les aspects juridiques a très vite été lié à une réflexion sur l'éthique du chercheur et l'occasion d'une démarche réflexive sur ses méthodes.
Dans un premier temps, le groupe de travail s'est orienté vers l'élaboration par la communauté scientifique "de bonnes pratiques" avec les contraintes suivantes : premièrement il n'existe pas de réponses juridiques simples à l'exploitation de l'oral et à la transcription des données et deuxièmement les solutions passent systématiquement par un travail réflexif sur la démarche du chercheur, seul moyen pour qualifier le statut des enregistrements et les objets exploités. Les "bonnes pratiques " consistent donc à clarifier les questions juridiques, mais aussi – et c'est là un point fondamental – à porter une réflexion sur le travail scientifique des linguistes dans le respect d'une éthique validée par la communauté scientifique.

---

[1] Art 5 de la Loi n°82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France, aujourd'hui art. L 111-1 du code de la recherche. *JO* du 16-07-1982, p. 2273 et ss.
[2] *Corpus oraux, Guide des bonnes pratiques op. cit.*, p. 36.
[3] Clapi-Icar http://clapi.univ-lyon2.fr
[4] EPML50 (ex Asila)
[5] Archivage du Lacito : http://lacito.vjf.cnrs.fr/archivage/index_fr.html
[6] PFC http://www.projet-pfc.net
[7] C-Oral-Rom 2005.

2. Aspects juridiques

D'une façon très schématique la réponse aux questions juridiques consiste à définir le statut juridique de l'objet "corpus" par ses conditions d'élaboration et sa composition, afin de procéder à la gestion contractuelle des droits des personnes concernées et de définir les responsabilités de ceux qui vont intervenir dans la vie du corpus (créateurs, hébergeurs, diffuseurs,…).

2.1. Définition de l'objet

Pour des raisons épistémologiques et techniques, la forme des corpus oraux est relativement complexe. Dans la majorité des cas les corpus oraux sont constitués :

- d'enregistrements (analogiques ou numériques) qui en cas de supports analogiques ont une durée de vie très courte avec une perte de qualité lors des migrations,
- de données contextuelles sur les locuteurs et la situation d'enquête qui peuvent être en partie des données personnelles (nom propre, profession, adresse, lieu, …),
- de transcriptions (sous la forme de fichiers indépendants ou permettant une synchronisation sur le signal ; transcription phonétique, orthographique, multilinéaire, etc.),
- d'annotations "secondaires" (informations sur les conditions de production des énoncés, précisions sur les phénomènes sonores tels que les rires et les bruits),
- d'annotations enrichies (étiquetage morphologique, syntaxique, annotations prosodiques pragmatiques, …),
- d'une documentation.

2.2. Domaines juridiques concernés

Pour définir le statut juridique de l'objet scientifique "corpus oral" et les droits des personnes concernées, il faut tout d'abord connaitre les conditions d'élaboration du corpus et de ses différentes composantes. Il s'agit ensuite de définir si le corpus est constitué d'informations du domaine public et/ou s'il est le produit d'une ou plusieurs créations intellectuelles susceptibles d'être protégées par le droit d'auteur. Il convient enfin de vérifier si le corpus contient des données personnelles qu'il faudra alors traiter. Ces statuts juridiques déterminés et les droits qui en découlent connus, il convient de s'enquérir des modalités de la gestion contractuelle de ces droits et de savoir si les titulaires de ceux-ci se sont prononcés sur les conditions de mise à disposition et de réutilisation des corpus – en apportant par exemple, leur consentement d'une manière formelle.

2.3. Diffusion scientifique et droit d'auteur

Seule une explicitation rigoureuse de la démarche du chercheur permet de savoir si un corpus est protégé par le droit d'auteur. Si tel est le cas, quels sont ces droits ?

Il convient de distinguer les droits patrimoniaux des prérogatives du droit moral. Les droits patrimoniaux se résument en un droit exclusif au profit de l'auteur (ou des titulaires) ou des ayants droit (bénéficiaires d'une cession, héritiers...) d'autoriser ou d'interdire la reproduction ou la communication au public de l'œuvre protégée. Quant aux prérogatives du droit moral, toujours attachées à la personne physique créatrice de l'œuvre protégée, elles sont au nombre de quatre : le droit de divulgation, le droit de repentir et de retrait, le droit à la paternité et le droit au respect de l'œuvre. En réalité, il existe une possibilité intermédiaire où les corpus protégés par le droit d'auteur peuvent être mis en libre accès dans le cadre d'une licence accordée par les titulaires de droits autorisant l'utilisation et l'exploitation des résultats (c'est le cas des Creatives commons). Sans être dans le domaine public, ces corpus sont – de par la volonté de leurs créateurs – libres d'accès et d'utilisation. Néanmoins, si les créateurs peuvent renoncer à exercer leurs droits patrimoniaux, il ne leur est pas possible de renoncer à leur droit moral qui reste imprescriptible.

3. Éléments pour de *bonnes pratiques*

3.1. Expliciter la démarche du chercheur

Les objectifs scientifiques, liés à la constitution, à l'exploitation, à la conservation et à la diffusion des corpus oraux sont très diversifiés, et le respect de ceux-ci, ainsi que leur hétérogénéité, impliquent que soit reconnue la diversité des démarches qui peuvent être adoptées par les chercheurs et par les utilisateurs ultérieurs de ces corpus.

Le Guide des bonnes pratiques n'a pas vocation à contraindre cette démarche en prescrivant une méthodologie type, mais souhaite fournir toutes les informations nécessaires au repérage des points juridiques et éthiques « sensibles ». Seule l'identification précise et détaillée des éléments

de la situation en jeu et notamment de la forme des données et de leurs supports, des pratiques de terrain, mais aussi des différentes étapes du traitement, permet d'apporter à la fois des éléments de réponses juridiques correspondant à la situation, et une évaluation des « risques » éventuels. Enfin, une analyse réflexive sur la démarche liée à la constitution et aux traitements des corpus oraux est le premier élément de l'élaboration d'une éthique reconnue par l'ensemble d'une communauté scientifique.

3.2. Le recueil de consentement

Le geste éthique le plus classique de la démarche du chercheur-enquêteur est le recueil de consentement du témoin. En réalité cette pratique est peu maîtrisée et souvent réduite à un formulaire de demande d'autorisation qui évoque en une phrase "le cadre d'un programme de recherche". Or sans informations préalables précises la demande d'autorisation n'a pas d'objet ni de sens. Pour que cette autorisation soit pertinente il conviendrait de concevoir le recueil d'un consentement "éclairé" qui démontre que le signataire est informé des finalités de la recherche et des conséquences à son égard d'une participation au projet.

Dans le cadre du recueil de données et notamment d'enregistrement pour des corpus oraux, le consentement devrait tenir compte de l'adéquation au destinataire (les informations fournies, pour être comprises doivent être adaptées aux compétences de compréhension du destinataire), et de l'explicitation des finalités de l'enquête (qui toutefois ne doivent pas renforcer le paradoxe de l'observateur en pointant l'objet de l'observation).

De plus, les explications sur le projet scientifique, doivent être complétées par des informations précises comme par exemple : les responsables de l'enquête et leur affiliation institutionnelle, ainsi que les financeurs ; une adresse de contact, les personnes qui auront accès aux données et qui travailleront sur elles, la façon dont les données seront anonymisées, le fait que les données seront transcrites selon des conventions particulières, la façon dont les données seront archivées une fois l'enquête terminée, les modalités d'accès aux informations relatives au projet et concernant tout particulièrement les données/analyses faisant référence à la personne (possibilité d'accès aux fichiers et informations concernant tout particulièrement la personne), les droits de la personne, notamment le droit de rétractation, les risques éventuels ainsi que les retombées positives, morales ou matérielles, de l'étude.

Enfin, le consentement devra préciser l'objet de la demande : les actions effectuées par les chercheurs dans le cadre du projet, les formats et les conditions de l'enregistrement, les conditions de diffusion des données et des résultats, les contextes de diffusion des données et des résultats.

Il est à noter que les formes de l'autorisation ne sont pas imposées par le législateur et qu'une demande orale enregistrée peut être valide et même parfois indispensable.

Sur le plan juridique, la collecte de données sensibles sans recueil de consentement est possible à la condition particulière que les données soient anonymisées dans un très bref délais. La procédure d'anonymisation est également très importante pour obtenir l'accord des témoins *a fortio*ri dans le cas d'une diffusion des données primaires.

3.3. L'anonymisation

Les pratiques actuelles des chercheurs en terme d'anonymisation se réduisent la plupart du temps à une opération de masquage d'un nom propre, d'une adresse ou d'un numéro de téléphone. Afin de vérifier la validité de ces pratiques et d'en définir les modalités, il convient de reposer avec précision la question légale qui est celle de l'impossibilité d'identifier des personnes. En effet, l'objectif est de protéger la vie privée des personnes enregistrées en dépersonnalisant les données, ce qui a amené le législateur à ne pas réduire cette identification à la simple présence de données nominatives.

Ainsi, si techniquement l'anonymisation consiste au remplacement ou au codage des données sensibles par des éléments neutres selon les supports concernés (remplacement par un blanc ou un pseudo à l'écrit, par un bip dans les fichiers sons et par floutage des visages sur les enregistrements vidéos), il serait erroné de penser que cette solution ne demande pas une expertise plus approfondie des risques d'exploitation d'éléments "dénommant".

3.4. Structure du corpus

Il existe d'autres possibilités que l'anonymisation par cryptage. Celles-ci reposent sur des limitations techniques prévues par la structure du corpus. La loi québécoise « concernant le cadre

juridique des technologies de l'information » propose de protéger l'anonymat non pas en modifiant les données, mais en limitant les possibilités de recherche, voire en les adaptant à la personne qui consulte la base selon des critères bien précis (sa profession, une autorisation, sa présence dans le fichier, etc.)

Cette dernière perspective offre pour la constitution et l'exploitation de corpus oraux la possibilité de faire coïncider les obligations légales avec les nécessités du travail de recherche. Toute donnée étant potentiellement sensible, une anonymisation systématique s'avère de plus en plus complexe ; elle peut même mettre en danger l'intérêt de certaines recherches. En effet, des détails concernant les personnes comme par exemple le nom, ou le lieu d'habitation peuvent constituer un élément important du corpus, ainsi que des résultats que l'on peut en tirer. C'est pourquoi la possibilité de ménager des niveaux d'accès selon des critères stricts (ex : chercheur ou non, présence d'autorisation, but de la consultation, etc.) semble une alternative efficace. Il existe d'autres procédés à inventer. En effet, l'article 11-2 de la nouvelle loi ouvre la possibilité de faire certifier des techniques nouvelles par la CNIL.

4. Conclusion

La démarche originale présentée ici a plusieurs intérêts. Outre le fait qu'elle offre les garanties d'une diffusion des corpus pour la recherche et pour d'autres finalités, elle impose une posture éthique aux collecteurs, utilisateurs et diffuseurs de corpus. C'est aussi l'occasion de porter un regard réflexif sur des pratiques et sur une démarche scientifique peu souvent explicitée. Enfin, il s'agit de permettre la constitution de corpus dont la mutualisation est la première étape d'une démarche scientifique rigoureuse qui ouvre les portes de l'analyse et de l'interprétation.

BIBLIOGRAPHIE

BAUDE, O. 2006. *Corpus oraux, Guides des bonnes pratiques, 2006*, CNRS-Editions et Presses Universitaires d'Orléans.

BAUDE, O. 2004. Les corpus oraux entre science et patrimoine. L'expérience de l'observatoire des pratiques linguistiques, in *Actes du Colloque international du GRESEC « La publicisation de la science »* (Grenoble), pp. 7-11.

BIBER, D. 1985. *Variations across spoken and written language*, Cambridge, CUP.

BIBER, D. 1999. *Longman Grammar of Spoken and Written English*, Londres, Longman.

BILGER, M. (dir.) 2000. *Linguistique sur corpus, études et réflexions*, Cahiers de l'université de Perpignan, Perpignan, Presses universitaires.

BILGER, M. (éd.) 2000. *Corpus, Méthodologie et applications linguistiques*, Paris, Champion.

BLANCHE-BENVENISTE, Cl. & JEANJEAN, C. 1987. *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

CALLU, A. & LEMOINE, H. 2004. *Patrimoine sonore et audiovisuel français : entre archive et témoignage : guide de recherche en sciences sociales*, 7 vol., 1 CD-Rom, 1 DVD-Rom, Paris, Belin.

CAMERON, D., FRAZER, E., HARVEY, P., RAMPTON, M. & RICHARDSON, K. 1991. *Researching Language : Issues of Power and Method*, London, Routledge.

CONDAMINES, A. (éd.) 2006. *Sémantique et corpus*, Paris, Hermès.

CRESTI, E. & MONEGLIA, M. (éds.) 2005. *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphie, Benjamins.

CRIBIER, F. & FELLER, E. 2003. *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile »* rapport présenté à Madame la Ministre déléguée à la Recherche et aux nouvelles technologies, dactylogr. 2 vol. et http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf

ENCREVE, P., & FORNEL de, M. 1983. Le sens en pratique, *ARSS* 46, L'usage de la parole.

HABERT, B., NAZARENKO, A. & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, A. Colin.

JACOBSON, M. 2004. Corpus oraux en linguistique de terrain, *Traitement Automatique des Langues*, 45/2, pp. 63-88.

JACOBSON, M. 2004. Les archives sonores au LACITO, *Bulletin de liaison de l'AFAS* 26 (http://afas.mmsh.univ-aix.fr/bulletin/Bulletin AFAS 26.pdf).

JOUTARD, P. 1979. Historiens, à vos micros. Le document oral, une nouvelle source pour l'histoire, *L'Histoire* 12, pp. 106-113.

KENNEDY, G. 1998. *An introduction to Corpus Linguistics*, Londres, Longman.

LABOV, W. 1972. *Sociolinguistic Patterns*, Philadelphie, University of Pennsylvania Press.

LEECH, G. 1992. The state of the art in corpus linguistics, Aijmer & Altenberg (éds.), pp. 8-29.

MONDADA, L. 1998. Technologies et interactions sur le terrain du linguiste. Le travail du chercheur sur le terrain. Questionner les pratiques, les méthodes, les techniques de l'enquête, Actes du Colloque de Lausanne 13-14.12.1998, *Cahiers de l'ILSL*, 10, pp. 39-68.

MONDADA, L. 2006. Video recording as the reflexive preservation-configuration of phenomenal features for analysis, in H. Knoblauch, J. Raab, H.-G. Soeffner, B. Schnettler (éds.).

MONDADA, L. (à paraitre) La demande d'autorisation comme moment structurant pour l'enregistrement et l'analyse des pratiques bilingues, *Tranel*, Université de Neuchâtel.

QUÉRÉ, L. *et al.* (éds.) 1984. *Arguments ethnométhodologiques*, Paris, Centre d'Étude des Mouvements Sociaux, EHESS.

Recherches sur le Français Parlé, 5, 1984. Pourquoi le français parlé est-il si peu étudié ?

Revue Française de Linguistique Appliquée, 1996. 1-2, 1999. IV-1.

SACKS, H. 1984. Notes on methodology, in J. M. Atkinson & J. Heritage (éds.), pp. 21-27.

SHAFFIR, W.B. & STEBBINS, R. A. (éds.) 1991. *Experiencing Fieldwork : An inside View of Qualitative Research*, Londres, Sage.

SILVERMAN, D. (éd.) 1997. *Qualitative Research. Theory Method and Practice*, Londres, Sage.

SINCLAIR, J. 1991. *Corpus, Concordance, Collocation*, Londres, OUP.

SINCLAIR, J. 1996. *Preliminary recommendations on corpus Typology*, Technical Report, Eagles.

SINCLAIR, J. & COULTHARD, R. M. 1975. *Towards an Analysis of Discourse*, Londres, OUP.

« Speech Annotation and Corpus Tools », A special issue of *Speech Communication* 33, 1-2 2001. Steven Bird and Jonathan Harrington.

WELLAND, T. & PUGSLEY, L. (éds.) 2002. *Ethical Dilemmas in Qualitative Research*, Aldershot, Ashgate.