HAL
open science

Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux

Olivier Baude

To cite this version:

Olivier Baude. Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux. Revue Française de Linguistique Appliquée, 2007, Corpus état des lieux et perspectives, XII (1), pp.85-98. halshs-01163043

HAL Id: halshs-01163043

https://shs.hal.science/halshs-01163043

Submitted on 11 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aspects juridiques et éthiques
de la conservation et de la diffusion des corpus oraux

Olivier Baude.
Université d'Orléans / DGLFLF

Résumé : *La numérisation des corpus de données sonores et multimodales ouvre de larges perspectives pour les sciences du langage. Toutefois, la conservation et l'exploitation de ces corpus oraux posent de nouveaux problèmes éthiques et juridiques que la communauté scientifique doit prendre en compte. Cet article présente les résultats d'un groupe de travail interdisciplinaire qui a rédigé un* Guide des bonnes pratiques *pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux.*

Abstract : *The digitalization of spoken language corpora opens large perspectives for linguistics. However, the archiving and the exploitation of these spoken corpora raise new ethical and legal problems that the scientific community must take into account. This article presents the results of an interdisciplinary working group which wrote a* Guide of good practices *for the constitution, the exploitation, the archiving and the diffusion of spoken language corpora.*

0. Introduction

Si depuis plus de 30 ans le domaine de la linguistique de corpus s'est considérablement développé autour des corpus écrits (masse de données disponibles, élaboration d'outils de traitement automatique de celles-ci), la situation est totalement différente pour les corpus oraux[1] et ce n'est que très récemment que les questions de conservation et de diffusion de ceux-ci se posent. Les toutes nouvelles technologies en matière de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils de traitement automatique du langage (transcriptions synchronisées sur le signal, annotations, etc.) ouvrent des perspectives prometteuses pour les études sur les corpus de langues parlées et devraient permettre de renouveler les sciences du langage en modifiant le champ de la linguistique de corpus. Toutefois cette situation ne va pas sans poser de nombreuses questions techniques, méthodologiques et théoriques mais aussi juridiques et éthiques. Ce sont principalement ces deux derniers aspects qui seront

[1] 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.

abordés dans cet article consacré au témoignage d'un travail interdisciplinaire qui s'est concrétisé par la rédaction d'un *Guide des bonnes pratiques* pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux[2]. Il s'agira donc, après avoir présenté le contexte - scientifique, épistémologique et politique - d'une telle initiative, de repérer concrètement les problèmes juridiques et éthiques afin d'éclairer la démarche des chercheurs et de proposer des solutions sous la forme d'un travail réflexif anticipant l'élaboration de bonnes pratiques.

1. Contextes

1.1. Objets émergents et émergence des problèmes

Il y a bien entendu des raisons épistémologiques propres au champ de la linguistique pour expliquer le retard pris en France pour la constitution et l'exploitation de corpus de données orales (Blanche-Benveniste 1987 ; Bergounioux 1992 ; Baude 2004). Cependant, et parallèlement à une modification du champ qui rend ses lettres de noblesse à la description de données attestées, le développement des nouvelles technologies facilitant la manipulation de masses de données sonores va transformer radicalement l'objet « données orales ».

En effet, si les outils informatiques permettent depuis peu de numériser et conserver en masse des enregistrements de la voix, de les diffuser facilement par les réseaux internet et autres, et de manipuler aisément des fichiers sons volumineux, les outils de reconnaissance et de synthèse de la parole mais aussi la transcription synchronisée sur le signal offrent aux corpus oraux l'accès aux outils de traitement automatique des corpus écrits créant ainsi un nouvel objet scientifique du fait même de cette synchronisation son/texte. Ce sont ces nouveaux objets, qui mêlent l'écrit et l'oral, qui ont fait émerger avec insistance des problèmes de conservation et de diffusion des corpus pour deux raisons principales.

- La présence de données primaires telle que la « voix » dans un corpus rappelle qu'il n'y a pas d'oralité sans « locuteur ». Il devient alors beaucoup plus difficile d'oublier que ce locuteur est une personne dont il convient pour des raisons tant juridiques qu'éthiques de respecter les droits.

- La linguistique de l'oral est constituée de différents domaines ; or si celui de la parole est méthodologiquement moins éloigné des corpus écrits, ce n'est pas le cas du français parlé, de la sociolinguistique, de la langue en interaction, de l'ethnolinguistique, etc., où le locuteur-témoin est particulièrement présent, par les données sociologiques souvent intégrées aux corpus, mais aussi par la réflexion sur les conditions de collecte des enregistrements (et sur l'impact de la diffusion des analyses pour la communauté observée).

---

[2] L'ouvrage *Corpus oraux, guide des bonnes pratiques 2006* est le résultat des travaux d'un groupe de réflexion réuni autour d'Isabelle de Lamberterie. Il a été rédigé par O. Baude coordinateur (Université d'Orléans/DGLFLF), C. Blanche-Benveniste (EPHE/Université de Provence), M-F. Calas (DMF), P. Cappeau (Université de Poitiers), P. Cordereix (BnF), L. Goury (CNRS), M. Jacobson (CNRS), I. de Lamberterie (CNRS) C. Marchello-Nizia (ILF/ENS-LSH-Lyon) et Lorenza Mondada, (Université Lyon 2).
Cet article est une présentation des travaux de ce groupe, la paternité de la grande majorité du contenu en revient donc à l'ensemble des auteurs à qui le rédacteur exprime toute sa gratitude.

1.2. Définition de l'objet : composition d'un corpus oral

Pour ces 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) et qui en cas de supports analogiques ont une durée de vie très courte avec une perte de qualité lors des migrations, de productions de locuteurs ;
- 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 primaires (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)

1.3. Cadres politiques de diffusion de la recherche

La complexité du contenu des corpus oraux et notamment la présence de données personnelles mais aussi le lourd investissement que représente la constitution comme le traitement (ne seraient-ce que les fastidieuses opérations de transcription) et l'enrichissement du corpus ont longtemps été invoqués pour expliquer l'absence de disponibilité et de diffusion de ceux-ci.

Récemment ces contraintes se sont confrontées aux cadres politiques de la diffusion de la recherche. Ainsi, depuis 1982 et la loi pour la recherche et le développement technologique en France[3], la diffusion des résultats fait partie des missions des chercheurs. De plus nous sommes dans une dynamique d'échange de l'information scientifique comme le confirme la déclaration de Berlin signée par la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) le 22 octobre 2003 sur *le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales*, dont l'objectif est de promouvoir Internet « comme instrument fonctionnel au service d'une base de connaissance globale de la pensée humaine » (*Corpus oraux*, 36). 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).

Cette dernière notion de standards d'interopérabilité se retrouvent dans différentes initiatives internationales (TEI, groupe de travail ISO TC37 SC4 pour la gestion des ressources linguistiques, protocole d'échange norme ANSI/NISO Z39.50, 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 se dotait en 2005 d'une direction de l'information scientifique et développait un an plus tard des centres de ressources numériques.

---

[3] 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, pp. 2273 et ss.

Conjointement à ce cadre politique, des laboratoires de recherche lançaient également différentes initiatives pour la diffusion et l'accessibilité des corpus oraux (Base CLAPI du laboratoire ICAR[4], projet Corpus Oraux de l'EPML 50[5], programme Archivage du LACITO[6], constitution de grands corpus disponibles comme par exemple, le projet Phonologie du Français contemporain[7]).

Face à cette volonté de conserver et diffuser des corpus oraux par des communautés scientifiques aux pratiques très différentes, de nombreuses questions notamment juridiques constituent un frein : Quelles autorisations doit-on faire signer à des locuteurs enregistrés dans le but de constituer un corpus de transcriptions exploitées et diffusées ? Est-ce qu'il faut que le corpus respecte l'anonymat ? A qui appartient le corpus ? Qui peut le vendre ? Qui est responsable du traitement et de la diffusion du corpus ? Quelles données (et métadonnées) peut-on conserver et diffuser sous la forme de fichiers informatiques ? Qui peut décider de la diffusion ou de la non diffusion des corpus ? Que doit on faire pour assurer la conservation des enregistrements (et qui peut assurer cette conservation sur du long terme) ? etc.
Les réponses à ces questions existent maintenant sous la forme d'un *Guide des bonnes pratiques* dont l'originalité de la démarche mérite d'être soulignée.

1.4. Une initiative d'élaboration de « bonnes pratiques »

Dans le cadre de son programme « corpus de la parole », la DGLFLF (Délégation Générale à la Langue Française et aux Langues de France, Ministère de la Culture) a créé un groupe de travail pluridisciplinaire regroupant des juristes, linguistes, conservateurs, informaticiens et représentants des fédérations de recherche en linguistique du CNRS. Ce groupe de travail s'est donné comme objectifs de recenser les pratiques actuelles et de définir en priorité les contraintes méthodologiques et théoriques liées à la recherche, de diffuser une synthèse sur la législation existante, d'établir des recommandations, et, le cas échéant, en cas de vide ou de flou, de formuler des propositions pour l'élaboration de normes et règles juridiques (notamment européennes) (*Corpus oraux*, 20).

Au fil du travail les propositions ont pris la forme d'un *Guide des bonnes pratiques pour la constitution, l'exploitation, la diffusion et la conservation des corpus oraux*.

2. Les aspects juridiques des corpus oraux en cinq questions[8]

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 définir les responsabilités de ceux qui vont intervenir dans la vie du corpus (créateurs, hébergeurs, diffuseurs…). Ces aspects juridiques peuvent se réduire à cinq questions essentielles.

Q1 : *Quel est le statut juridique de l'objet « corpus oral » ?*

---

[4] CLAPI-ICAR http://clapi.univ-lyon2.fr.
[5] EPML50 (ex Asila).
[6] Archivage du LACITO : http://lacito.vjf.cnrs.fr/archivage/index_fr.html
[7] PFC http://www.projet-pfc.net
[8] Cette partie de l'article reprend les éléments principaux rédigés par I. de Lamberterie pour le *Guide*.

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.

Q2 : *Qui est le propriétaire d'un corpus ?*

Le statut juridique d'un corpus dépend des données qu'il contient. Soit il dépend du domaine public, et est alors libre d'exploitation pour tout le monde, soit il appartient à un *auteur*, et est alors protégé.

Le domaine public recouvre non seulement les idées de liberté d'accès et de gratuité d'utilisation des données, mais aussi la possibilité pour chacun de les exploiter. Il se caractérise, en outre, par l'absence de monopole puisque les informations qui tombent dans le domaine public deviennent de facto des « choses communes » (*Corpus oraux*, 38).

Bien entendu la distinction n'est pas définitive, ainsi les œuvres protégées par le droit de la propriété intellectuelle, notamment par le droit d'auteur ou les brevets, finissent par entrer dans le domaine public. En France, c'est 70 ans après la mort de son auteur qu'une œuvre tombe dans le domaine public, même si à l'expiration de ce délai, d'autres types de protection peuvent subsister sur les œuvres de l'esprit : les droits patrimoniaux d'une part, les attributs imprescriptibles du droit moral d'autre part. En conséquence certaines œuvres (ou éléments de ces œuvres) tombées dans le domaine public, peuvent encore être protégées par les dispositions qui concernent le droit moral.

Les corpus oraux, ou tout au moins les enregistrements et les transcriptions relèvent-ils des considérations évoquées ? Comme nous allons le voir, la réponse n'est pas simple et les enregistrements linguistiques suscitent ainsi de nombreuses hésitations. Il est en effet complexe de savoir si le contenu d'une langue, son expression phonique, font ou non partie du domaine public. En outre, ce fonds commun est-il universel ou bien seulement commun à une petite communauté ? Aujourd'hui, il fait de plus en plus l'objet de revendications identitaires qui soulèvent de nouvelles interrogations (*Corpus Oraux*, 38).

La protection du droit d'auteur demande également une définition précise des objets et des opérations de traitement de ces objets. En effet pour qu'un corpus soit protégé par le droit d'auteur il faut que trois conditions soient remplies:

Il faut en premier lieu qu'il corresponde à l'exigence d'une activité créatrice : un travail de compilation d'informations n'est pas protégé en soi. Pour être protégé, il est par ailleurs indispensable que le corpus ait une forme définie. Ce qui est protégé ce n'est pas le contenu du corpus mais son enveloppe, son architecture. Enfin, la forme du corpus doit répondre à la condition d'être originale. L'auteur est en principe la (ou les) personne(s) physique(s) sous le nom de laquelle (ou desquelles)

l'œuvre est divulguée. Le travail scientifique suppose l'intervention de nombreux acteurs dont bon nombre sont susceptibles de revendiquer la qualité d'auteur sur les résultats de la recherche. Certains corpus oraux, comme les autres produits de la recherche, peuvent rester l'œuvre d'un auteur unique, alors que d'autres peuvent être l'œuvre de plusieurs auteurs. Dans le cas de pluralité d'auteurs, le droit distingue les œuvres de collaboration des œuvres collectives. Pour les premières, chaque co-auteur dispose des mêmes prérogatives. D'autres œuvres - telles que les bases de données ou les dictionnaires - peuvent être qualifiées d'œuvre collective lorsqu'elles sont créées « sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et sous son nom et dans laquelle la contribution personnelle des divers auteurs ... se fond dans l'ensemble » (Art. L 113-2 du CPI). Dans ce dernier cas, c'est la personne physique ou morale qui a pris l'initiative de l'œuvre qui dispose des droits d'auteur.

Par ailleurs, le contexte de la création ou le statut de l'auteur peuvent avoir des incidences sur la détermination du titulaire des droits d'auteur. L'œuvre a-t-elle été créée dans le cadre d'une mission de service par un employé ou un fonctionnaire ? Quels sont les droits respectifs de l'auteur et de son employeur ? Si la question est résolue le plus souvent par le contrat de travail, elle reste plus délicate quand le créateur est un fonctionnaire.

Q3 : *Quels droits pour l'auteur d'un corpus ?*

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 droits (bénéficiaires d'une cession, héritiers...) d'autoriser ou 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. 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.

Q4 : *Qui est responsable (et de quoi) ?*

Tout traitement de données doit avoir un responsable, sa mission est d'éviter ou de circonvenir les risques inhérents à la gestion et l'utilisation des données recueillies. La loi lui fixe donc des obligations. La directive européenne 95/46/CE (« Flux de données transfrontières ») dans son article 2, repris pour la refonte de la loi « Informatique et libertés » donne la définition suivante : « *Le responsable d'un traitement de données à caractère personnel est, sauf désignation expresse par les dispositions législatives ou réglementaires relatives à ce traitement, la personne, l'autorité publique, le service ou l'organisme qui détermine ses finalités et ses moyens* ». Le responsable du traitement se doit de veiller à la qualité des données (adéquate, pertinente, non excessive et exacte, ce qui implique un droit d'accès, d'opposition et de rectification ouvert aux personnes concernées), au respect de leur(s) finalité(s) annoncées au préalable et au respect du principe de licéité (les

données doivent avoir été recueillies loyalement, ce qui implique le recueil d'un consentement éclairé). Il est, de plus, responsable de la déclaration à la CNIL, du recueil du consentement et du respect des règles de conservation et notamment de la confidentialité des données.

Q5 : *Comment se conformer au respect de la vie privée ?*

Si un corpus contient des données personnelles il dépend de la loi « Informatique et liberté ». La création d'un corpus passe le plus souvent par la collecte de données. Celles-ci pouvant être des données personnelles, cette collecte doit être faite dans le respect de la loi « Informatique et libertés » : licéité et loyauté, information préalable, obtention du consentement des personnes concernées, respect des finalités annoncées[9]… Si les données sont « anonymisées » de manière irréversible, elles sortent du champ de la loi et peuvent être conservées (voir annexe anonymisation). Toutefois dans la recherche, le besoin de « traçabilité » nécessite souvent de sauvegarder les données personnelles. L'anonymisation ne consiste pas simplement à faire disparaître les noms propres mais plutôt à gérer les données personnelles afin de ne pas permettre l'identification des locuteurs. Ainsi, La CNIL préfère parler de données personnelles : « *...pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne* ».

De fait l'identification d'une personne peut se faire de manière directe ou indirecte.
Le caractère personnel d'une donnée dépend des moyens de tri, de rapprochement qui pourraient être mis en œuvre. Cela conduit donc à une évolution constante du champ des données personnelles, la technique mettant à la disposition du plus grand nombre des outils de requêtes plus en plus performants.

Certains auteurs avancent qu'il suffit qu'il y ait une probabilité suffisante de rapprochement à une personne pour qu'une donnée acquière un caractère personnel indirect. Les analyses ne sont pas toujours explicites mais il n'est possible de négliger cet argument (Lamy, *Droit de l'informatique et des Réseaux*, 508). Dans le domaine statistique, la CNIL a imposé des seuils au-delà desquels des rapprochements d'agrégats de données – pourtant individuellement anonymes – sont interdits.

Le caractère personnel d'une information dépend de l'objet qu'elle décrit, du contexte dont elle provient, mais aussi de la personne qui la reçoit. Pour pouvoir identifier un individu ou un groupe, nous avons besoin des informations, mais nous n'y arriverons pas sans un élément de connaissance propre qui déclenchera le mécanisme d'association. Lorsqu'on réfléchit à l'anonymisation, il convient donc de connaitre les éléments à traiter, mais aussi les opérations que vont subir les données (pour les corpus oraux : les données primaires (signal), les métadonnées, la transcription, l'annotation, le balisage, la diffusion, la conservation, etc.).

3. Eléments de réponse

[9] http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17_definitive-annotee.pdf

Les réponses à ces cinq questions sont loin d'être évidentes et la gestion des points juridiques et éthiques sensibles ne peut se restreindre à l'application de contraintes mais doit respecter les pratiques des chercheurs. Les objectifs, notamment scientifiques, liés à la constitution, à l'exploitation, à la conservation et à la diffusion des corpus oraux, sont très divers, et le respect de ceux-ci, ainsi que leur hétérogénéité, impliquent que soit reconnue la diversité des *pratiques* qui peuvent être adoptées par les chercheurs et par les responsables de la diffusion et de la conservation de ces corpus. Seule l'identification précise et détaillée des éléments de la situation en jeu (forme des données et de leurs supports, pratiques de terrain, étapes de leurs traitements), permettent 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.

3.1. Expliciter la démarche du chercheur

Comme nous l'avons évoqué plus haut, la définition du statut juridique d'un corpus ne peut se faire qu'après avoir explicité les éléments qu'il contient ainsi que les conditions de son élaboration et de son exploitation. Seul ce travail d'explicitation permet d'anticiper les problèmes juridiques et éthiques de conservation et de diffusion. La grille suivante permet de conduire ce travail en sept étapes descriptives : type de données, techniques d'enquêtes, rôles des participants et modes d'approche, lieu de l'enquête, dispositif d'enregistrement, annotations, exploitation-diffusion.

De quels types sont les données composant un corpus oral ?

On peut grossièrement opérer une dichotomie entre données primaires (enregistrements audio et vidéo mais aussi documents lus ou écrits durant l'action enregistrée) et données secondaires (annotations (et en particulier descriptions) des données primaires (métadonnées)). Cette distinction a deux utilités : caractériser la source et donc un état original (de base) du corpus et définir les niveaux du travail scientifique du chercheur (collecte, transcription, annotation, analyses,..).

Les données sont enregistrées et conservées sur des supports (physiques, magnétiques, optiques) selon un mode analogique ou numérique. La description du support et du mode est importante pour évaluer les possibilités de conservation et leurs conséquences sur l'objet « original ». Enfin l'explicitation de la structure des données et donc des codages, des formats et des formalismes utilisés permet de définir si un corpus répond à une technologie explicite, acceptée par une communauté et/ou ayant fait l'objet d'une normalisation et qui peut être ouverte à la communauté ou posséder un caractère propriétaire (brevets ou formats propriétaires).

Comment ces données ont-elles été collectées, quelles ont été les conditions d'élaboration du corpus ?

Pour répondre à cette question il convient d'abord de définir les techniques d'enquêtes utilisées : l'enregistrement en laboratoire selon un protocole expérimental produit des données et une situation aux caractéristiques juridiques bien différentes de l'enregistrement de récits de vie, de contes, de conversations professionnelles,

d'émissions radio ou télédiffusées, d'entretiens etc. Il faut ensuite caractériser les relations entre les enquêteurs et les enquêtés qui dépendent généralement du mode d'approche du chercheur (le témoin est-il rémunéré ? a-t-il exprimé son consentement ? est-il captif ?) et du rôle des participants ainsi que du lieu de l'enregistrement (est-ce un lieu public ou privé?). Enfin, il s'agit de décrire le dispositif d'enregistrement (audio ou vidéo, voyant ou caché, géré par le chercheur ou par les témoins, etc.) qui matérialise très souvent les aspects juridiques et éthiques de la relation observateur-observé.

Quels sont les traitements subis par les données primaires collectées ?

L'exploitation par les chercheurs des enregistrements oraux et multimodaux modifient considérablement l'objet. Ainsi les opérations d'annotations et notamment la transcription intègrent de nombreux aspects théoriques et interprétatifs mais aussi juridiques et éthiques. De fait,

> *dans le passage de l'oral à l'écrit graphico-visuel, de nombreuses opérations de catégorisation sont effectuées, soit quant aux formes linguistiques, segmentées visuellement en unités (Blanche-Benveniste & Jeanjean, 1987 ; Mondada, 2000), soit quant à l'identité des locuteurs eux-mêmes (Mondada, 2003). Du point de vue de la protection de l'image et de l'identité des personnes enquêtées et enregistrées, il convient d'apprécier ces effets pour éviter la surinterprétation, la stéréotypisation (Jefferson 1996) et la stigmatisation des locuteurs et de leurs façons de parler* (*Corpus oraux*, 73).

Les opérations d'annotations ne sont donc jamais neutres en termes d'effet de la recherche sur des données produites par les locuteurs. La description rigoureuse des traitements permet d'évaluer les effets de ceux-ci sur des productions dont on doit respecter les droits de leurs auteurs. De plus cette description fournit la possibilité d'attribuer la paternité et la responsabilité de niveaux de traitement à un ou plusieurs auteurs.

Quelles sont les pratiques de conservation et de diffusion ?

L'explicitation de la démarche de conservation et de diffusion revient à déterminer ce qui sera accessible et les modalités de cet accès (qui, quand, comment?).

L'explicitation de la démarche du chercheur permet donc de définir les données qui composent un corpus ainsi que les conditions d'élaboration de celui-ci. Ce travail fait, il est beaucoup plus aisé de repérer les questions juridiques qui se posent dans les deux grands domaines juridiques concernés : les droits d'auteur (moraux et patrimoniaux) et le respect de la vie privée. Cette description de la démarche est une première étape dans l'élaboration de bonnes pratiques qui doit logiquement être complétée par deux aspects qui semblent indispensables pour le respect de l'éthique du chercheur et qui permettront d'anticiper les problèmes de conservation et de diffusion. Ces deux aspects sont le recueil du consentement de l'enquêté et les procédures d'anonymisation des données.

3.2. Les bonnes pratiques pour un consentement éclairé

Pour des raisons éthiques, les chercheurs collectant des enregistrements produits par des informateurs souhaitent obtenir leur consentement. Dans le cas où le corpus

contient des données protégées par le droit d'auteur ou étant considérées comme des *données personnelles*[10] cette collecte doit obligatoirement être faite dans le respect de la loi « Informatique et libertés » et donc respecter les notions de licéité et loyauté, d'information préalable, du respect des finalités annoncées et a fortiori d'obtention du consentement des personnes concernées [11].

Le recueil de consentement n'est pas une opération banale dans les pratiques des chercheurs, parfois inexistante 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

---

[10] Selon la Loi du 6 aout 2004, « Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne ».

[11] http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17_definitive-annotee.pdf

très bref délais. La procédure d'anonymisation est également très importante pour obtenir l'accord des témoins a fortiori dans le cas d'une diffusion des données primaires.

3.3. Les bonnes pratiques de l'anonymisation

Les pratiques actuelles des chercheurs en termes 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 présence de données nominatives. Ainsi, les textes législatifs abandonnent les termes de « données nominatives » au profit de « données personnelles » :

> *...pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne (directive 95/46/CE).*

Une attention toute particulière doit donc être apportée aux données personnelles contenues dans un corpus qui permettraient d'identifier directement un témoin (formes nominatives, données personnelles, profession, statut, titres, activités sociales, parenté, réseaux, référence à des lieux, référence à des caractéristiques de la personne, caractéristiques physiques, etc.) mais aussi à tout ce qui peut permettre indirectement une identification (notamment les possibilités de recoupement d'informations).

Les procédures d'anonymisation concernent les données premières audio et/ou vidéo, les données premières textuelles, les données secondaires et les données secondaires visuelles en évitant une anonymisation sur les données premières originales et en gardant éventuellement la possibilité de travailler sur des données partiellement anonymisées dans un cadre de recherche précis même si en revanche, la diffusion implique une anonymisation rigoureuse.

Techniquement l'anonymisation consiste principalement 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).

Cependant le codage des données personnelles n'est pas la seule procédure d'anonymisation. La possibilité d'élaborer des bases de données séparées est utilisée en France. Plus intéressant encore « *les limitations techniques* » qui permettent de protéger l'anonymat sans modifier les données mais en limitant les possibilités de recherche et de croisement des données (ne pas permettre de croiser le lieu d'habitation et la profession par exemple). Cette dernière possibilité est particulièrement intéressante dans le cadre de recherche sur la langue parlée où les informations sociologiques pertinentes sont indispensables.

3.4. Des corpus oraux au patrimoine sonore, l'enjeu d'une normalisation

Les enregistrements de la voix, à la base des corpus oraux, ont été constitués en fonds sonores depuis presque un siècle (Calas 1996). La numérisation de ces fonds

sonores et l'informatisation des catalogues n'ont fait qu'accroître les relations entre les corpus oraux - objets scientifiques - et les archives orales - objets patrimoniaux. Ainsi plusieurs projets actuels de grands corpus oraux (Phonologie du Français Contemporain, Archivage du LACITO, Eslo) ont, en vue du dépôt de leurs données, sollicité les institutions de conservation qui elles mêmes développent la diffusion en ligne de leurs archives (Gallica pour la BnF, archives pour tous de l'INA). Le transfert de la gestion de la conservation et de la diffusion des corpus à une institution fait apparaître les mêmes problèmes que rencontrent les chercheurs : principe de cohérence des fonds, gestion des métadonnées, des formats et des normes, questions de propriétés intellectuelles et de protection de la vie privée objectifs de. Or si ces problèmes, qui sont nouveaux pour les chercheurs, n'ont pas été résolus par les institutions de conservation, c'est aussi parce que la question du statut de l'oral comme objet patrimonial est indissociable du statut de l'oral dans le champ scientifique.

En ce sens, les documents numériques et les technologies de conservation et de diffusion en masse posent d'une façon cruciale le problème de la normalisation des données (supports, formats, codages etc.) pour des usages diversifiés.

Les bonnes pratiques et la normalisation des données (*Corpus oraux*, 79-95)

La réponse technique à un problème méthodologique et éthique démontre l'importance des relations entre l'éthique, le théorique et le technique dans les questions d'exploitation de corpus. Les objectifs de partage de données, d'échange et d'interopérabilité sont ainsi indissociables des aspects de diffusion et de conservation des corpus et posent de fait la question des pratiques de normalisation même s'il ne peut y avoir de travail sur la normalisation des corpus sans prise en compte de l'impact de celle-ci sur les analyses. Il suffit pour s'en persuader d'évaluer l'enjeu de la normalisation pour un locuteur ou pour une communauté lié à la normalisation provoquée par les opérations de transcription.

4. Conclusion

Les objectifs de conservation et de diffusion des corpus ne reposent pas simplement sur des questions techniques et stratégiques de normalisation (pour permettre l'échange, la multifinalité et l'interopérabilité), il s'agit également d'élaborer des bonnes pratiques dans le respect de l'éthique et du juridique qui obligent le chercheur à *savoir ce qu'il fait* et la communauté scientifique à *prendre en charge les objets scientifiques qu'elle construit*. Ce n'est alors pas étonnant de découvrir que les questions juridiques se concentrent sur deux domaines : la propriété des données et le respect de la vie privée. Les réponses ne sont pas simples et ne sont pas systématiques, les contraintes de la recherche font que certains corpus ne peuvent à la fois suivre des bonnes pratiques respectant l'observé et des bonnes pratiques de diffusion des données. Il est fondamental de respecter l'hétérogénéité des pratiques de la recherche fondée sur les corpus oraux et de considérer que certains corpus ne sont pas fait pour être diffusés ni même accessibles en dehors de la relation enquêteur-enquêté. Cependant pour la grande majorité des cas, l'élaboration de bonnes pratiques par la communauté scientifique doit permettre de donner un véritable statut aux corpus oraux en imposant une démarche éthique qui reconnaît les droits des observés, en facilitant l'exploitation, la diffusion et la conservation par

une réflexion sur la normalisation des données et métadonnées et en anticipant les changements de finalités. Pour atteindre cet objectif, il revient au chercheur d'expliciter sa démarche, de la collecte à la diffusion, de gérer les droits repérés (de propriété intellectuelle et de respect de la vie privée) et de structurer les corpus pour faciliter l'interopérabilité et le cumulatif tout en permettant par une limitation technique de protéger l'accès aux différentes données.

Ainsi, l'élaboration de *bonnes pratiques*, loin d'être restreintes à l'application de contraintes juridiques est aussi l'opportunité pour une communauté scientifique de s'acquitter de la dette que le chercheur contracte envers son terrain, en offrant un véritable statut à son objet d'étude.

Olivier Baude
Université d'Orléans et DGLFLF
UFR LLSH, 10 rue de Tours, 45072 Orléans.
<olivier.baude@univ-orleans.fr>

Références

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), 7-11.

Becker, H.S. & Geer, B. (1960). Participant observation : the analysis of qualitative field data. In Adams & Preiss (eds.), 267-289.

Bergounioux, G. (ed.) (1992). Enquêtes, Corpus et Témoins. *Langue Française* 93.

Biber, D. (1999) *Longman Grammar of Spoken and Written English*. Londres, Longman.

Bilger, M. (ed.) (2000). *Corpus, Méthodologie et applications linguistiques*. Paris, Champion.

Blanche-Benveniste, C. (1997). Transcription et technologie. *Recherches sur le Français Parlé* 14, 87-100.

Bourdieu, P. (1993). *La misère du monde*. Paris, Le Seuil.

Calas, M-F. & Fontaine, J-M (1996). *La conservation des documents sonores*. Paris, CNRS Editions.

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

Condamines, A. (ed.) (2006). *Sémantique et corpus*. Paris, Hermes.

Cresti, E. & Moneglia, M. (eds.) (2005). *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam, Benjamins.

Cribier, F. & Feller, E. (2003). *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile »*. Rapport au Ministre délégué à la Recherche et aux nouvelles technologies, dactylogr., 2 vol. 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.

Gumperz, J.J., & Hymes, D. (eds.) (1972). *Directions in Sociolinguistics : The Ethnography of Communication*. New-York, Hold, Rinehart & Winston.

Habert, B., Nazarenko, A. & Salem, A. (1997). Les linguistiques de corpus. Paris, A. Colin.

Habert, B., (2005*), Instruments et ressources électroniques pour le français*, Ophrys, Paris.

Jacobson, M. (2004). Corpus oraux en linguistique de terrain. *Traitement Automatique des Langues*, 45/2, 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).

Kennedy, G. (1998). *An introduction to Corpus Linguistics*. Londres, Longman.

LAMY, Droit de l'informatique et des réseaux (S. Marcellin, L. Costes & *al.* eds., Paris, 2004).

Leech, G. (1992). The state of the art in corpus linguistics. In Aijmer & Altenberg (eds.), 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, 39-68.

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

Quéré, L. & *al.* ed. (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) I-2, (1999) IV-1.

Sacks, H. (1984). Notes on methodology. In J.M. Atkinson & J. Heritage (eds.), 21-27.

Sankoff, D., Sankoff, G., Laberge, S. & Topham, M. (1976). Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale. *Cahiers de Linguistique* 6, 85-125.

Silverman, D. (ed.) (1997). *Qualitative Research. Theory Method and Practice*. Londres, Sage.

Sinclair, J. (1996). *Preliminary recommendations on corpus Typology*. Technical Report, Eagles.

Speech Communication (2001) Speech Annotation and Corpus Tools. Vol. 33, 1-2, S. Bird & J. Harrington (eds.).

Welland, T. & Pugsley, L. (eds.) (2002). *Ethical Dilemmas in Qualitative Research*. Aldershot, Ashgate.