QUELQUES PISTES ET JALONS POUR UNE APPROCHE REFLEXIVE DE L'ELABORATION DES CORPUS EN ANALYSE DES DISCOURS ET DE LA COMMUNICATION PUBLICS

Jérôme JACQUIN
Université de Lausanne – CLSL[1]
jerome.jacquin@unil.ch

Résumé

Cette contribution entend problématiser la question de l'élaboration des corpus en analyse des discours et de la communication publics. En lieu et place d'une discussion théorique du problème, je propose d'aborder la question d'une manière à la fois pratique (par la description relativement factuelle d'une expérience concrète) et réflexive (par la problématisation des étapes de constitution et donc des choix et contraintes que cette pratique implique). Le corpus et les données apparaîtront ainsi comme des construits que le chercheur élabore de manière à assurer leur fonction de transparence (leur capacité à rendre compte d'un « réel »). Il s'agira par conséquent de penser les dimensions (corpus projeté, corpus réalisé et corpus consolidé) et les étapes (principalement le recueil et la confection des données) impliquées par cette élaboration.

Mots-clés : Corpus, données, méthodologie, recueil, confection, communication publique

1. Introduction

Cette contribution se donne comme objectif de filer l'élaboration d'un corpus spécifique en analyse des discours et de la communication publics. Plutôt que partir de considérations épistémologiques générales, elle entend en raconter l'histoire, c'est-à-dire – par l'adoption d'un regard réflexif sur une pratique de recherche effective – faire émerger les étapes, paramètres et problèmes significatifs de l'élaboration d'un tel corpus. Il s'agira donc de mener deux réflexions en parallèle : (*i*) décrire, le plus factuellement possible, les différentes étapes de constitution d'un corpus spécifique au sein d'une recherche particulière, tout en (*ii*) extrayant de la spécificité de la démarche quelques éléments généraux qui puissent servir aux étudiants inscrits au programme de spécialisation.

Dans un premier temps, je définirai rapidement ce que nous entendons par données et par corpus (2.). Si les définitions prendront progressivement chair au

[1] Centre de linguistique et des sciences du langage.

fil du texte, il s'agit surtout de proposer un point de vue identifiable et tranchant conduisant les étudiants à interroger leurs préconstruits. J'adopterai, après bien d'autres, une perspective résolument praxéologique du « recueil » des données et de la constitution des corpus. Tant les premières que les seconds sont des construits et ne prennent sens que relativement à une pratique et un objectif de recherche. Le corpus, en effet, c'est toujours le corpus de l'analyste. Ce point de vue nous invitera à aborder successivement trois dimensions du corpus : un corpus est projeté (en fonction d'objectifs de recherche), un corpus est réalisé (par le biais d'un travail de terrain), un corpus est consolidé (en vue d'une gamme d'applications). Le corpus étant fonction de l'objectif de recherche que se donne l'analyste, il s'agira donc de dire deux mots de ceux que l'analyse des discours et de la communication publics poursuit, d'une part, et qui ont guidé la réalisation de mon corpus, d'autre part (3.). Il s'agira ensuite d'en présenter les principales étapes de réalisation (4.). Si la nature et l'importance de ces étapes peuvent varier selon les corpus, tous nécessitent d'une manière ou d'une autre un travail de terrain et méritent donc d'être réfléchis comme tel. Nous verrons que le corpus projeté par l'objectif de recherche entre en relation dialectique avec celui que l'on réalise et que l'on consolide pour l'analyse. Avant de conclure (5.), il sera donc finalement et rapidement question du troisième état que le corpus est amené à embrasser : l'état « consolidé », c'est-à-dire un état synchronique, photographié à toutes fins utiles[2]. Si projection et réalisation dialoguent et s'enrichissent mutuellement, il arrive régulièrement – et cette contribution en est un exemple – que le chercheur doive consolider son corpus, c'est-à-dire en dessiner les frontières et être en mesure, grâce à des annotations métatextuelles, d'expliciter les liens qu'il tisse entre les données qui constituent son corpus.

2. Quelle définition des données et du corpus ?

2.1. Un point de vue radicalement praxéologique

Avant toute chose, il est nécessaire d'abandonner deux préconstruits que les termes « données » et « corpus » ont respectivement tendance à véhiculer :

[2] Ce point étant propre à chaque corpus, il ne fera pas l'objet d'un développement spécifique.

a) Tout d'abord, parler de « données » est un abus de langage[3]. Ce que nous analysons, en tant qu'analystes de la communication publique, ne nous est pas donné, loin de là. C'est à l'analyste de « se donner » les moyens de faire son analyse. Et pour ce faire, il construit ses données. Ne serait-ce que parce qu'il sélectionne une partie du flux des expériences, qu'il choisit, qu'il écarte, qu'il uniformise. Et cette justification du découpage, justification prenant la forme d'annotations qui ré-articulent les données au monde phénoménologiquement adjacent, est un travail de tous les instants, quelque chose qui se construit aussi pendant l'analyse, et qui est un des buts de l'analyse. L'analyste construit ses données *pour* leur octroyer une certaine forme de transparence, pour qu'elles puissent rendre un phénomène manifeste, accessible, descriptible. Ainsi, les données instancient les points de vue qui ont présidé à leur recueil (Berthelot 2000 : 161), elles sont fonction d'un objectif de recherche formulé dans le cadre d'une tradition disciplinaire.

b) Ensuite, un « corpus » n'est pas une collection de données, quelle que soit la manière dont on définit ces « données ». Un corpus n'est ni un « réservoir d'attestations », ni une « archive » (Rastier 2004). Il ne constitue pas la base de l'analyse, il n'en constitue même pas, nous le verrons, un préalable. Un corpus est un outil. Il est à la fois ce qui permet l'analyse et son horizon, dans un dialogue permanent entre l'analyste et les contextes dans lesquels il est mobilisé. Il accompagne le chercheur et sa recherche.

Ce sont ces deux idées – les données sont des constructions à prétention de transparence et le corpus est un outil plutôt qu'une ressource – qui guideront la présentation de l'élaboration de mon corpus. En bref, on admettra que tant la

---

3 Si la sociologie a dès ses fondements réfléchi à la relation qu'elle entretenait avec les « données » et plus généralement à la responsabilité du sociologue dans la construction des « faits sociaux » qu'il entend analyser (pour une introduction à cette problématique, voir par exemple Berthelot 2000 : 159 et ss. ; Bourdieu, Chamboredon & Passeron 2005 : 51 et ss.), cette réflexion est bien plus jeune en linguistique et plus globalement en sciences du langage et de la communication : c'est le développement de la *linguistique de corpus* dès la seconde moitié des années 1990 – on entend traditionnellement par « linguistique de corpus » l'analyse linguistique assistée, voire même réalisée, par ordinateur (Biber, Conrad & Reppen 1998 : 4) – qui a favorisé l'émergence d'une réflexion plus générale sur les rapports que l'analyste du langage entretient avec ses données et son corpus (voir par exemple Mayaffre 2002 & 2005 ; Rastier 2004 ; Mondada 2005). Deux problématiques complémentaires restent néanmoins encore peu questionnées : (*i*) l'opportunité et l'organisation de corpus hétérogènes, articulant textes écrits et documents audio-visuels ; (*ii*) la spécificité des corpus en analyse de la communication publique et des interactions médiatiques. La présente contribution entend donner quelques pistes de réflexion relatives à ces deux enjeux.

construction des données que la systématisation de ces données en corpus sont fonction d'un objectif de recherche, lui-même dépendant d'une tradition disciplinaire.

2.2. Différence entre corpus projeté et corpus réalisé

L'importance donnée à la définition d'un objectif de recherche implique de distinguer deux stades (ou dimensions) dans l'élaboration d'un corpus : d'une part, la formulation d'un objectif projette les contours d'un corpus en devenir (le *corpus projeté*, le corpus *en puissance*), d'autre part, la pratique du terrain sédimente progressivement un corpus effectif (le *corpus réalisé*, le corpus *en acte*). Cette dichotomie ne signifie pas pour autant une frontière hermétique, mais permet au contraire de penser une dialectique. Le corpus étant traité comme outil évolutif, il est l'objet d'une constante réflexion, projection et réalisation s'influençant réciproquement. Le corpus projeté n'est donc pas supposé opérer une détermination unilatérale sur le corpus réalisé. Les cas ne sont pas rares, en effet, où la réalisation progressive du corpus vient nuancer sa projection et même modifier les objectifs de recherche et leur formulation.

2.3. Différence entre corpus réalisé et corpus consolidé

Du fait de la dialectique qui s'opère avec l'objectif de recherche, l'élaboration du corpus implique la présence de phases de consolidation. Périodiquement, le chercheur doit – ne serait-ce que pour être en mesure d'en parler et de l'analyser – verrouiller son corpus, en expliciter la cohérence, l'histoire, les objectifs. Le corpus passe alors d'un stade *réalisé* à un stade *consolidé*. Toujours précaire, le stade de consolidation, dont le présent texte est un exemple, apparaît comme le moment où la dynamique générative [corpus projeté / corpus réalisé] est suspendue, le temps d'utiliser le corpus pour « faire autre chose ». Ce qui importe à ce stade, c'est que le corpus soit « solide », qu'il soit argumentable, qu'il puisse rendre manifeste sa logique.

Dans la suite de cette contribution, je vais considérer successivement les deux premières dimensions susmentionnées. Dans un premier temps, j'aborderai la question des objectifs de recherche et de la projection du corpus (3.). Dans un second temps, je passerai en revue les dimensions pertinentes d'un travail de terrain et par conséquent de la réalisation du corpus (4.). Étape de présentation et

d'analyse propre à chaque travail de recherche, la phase de consolidation ne sera qu'évoquée avant la conclusion (5.).

3. Objectif de recherche et corpus projeté

3.1. Tradition de recherche

La présente contribution s'inscrit dans un programme d'enseignement et de recherche en analyse des discours et de la communication publics. L'objectif de recherche qui oriente l'élaboration de mon corpus et qui en dessine par conséquent les contours est donc tributaire de son inscription dans un paradigme disciplinaire. Ceci est crucial. Lorsque l'on parle d'objectifs de recherche, on ne doit pas oublier que ceux-ci s'inspirent toujours d'une pratique existante. Cette pratique, consiste non seulement dans les façons de faire de la communauté de chercheurs dans laquelle on s'inscrit, mais aussi dans les valeurs à l'aune desquelles cette communauté évalue le caractère approprié, ou non, des corpus projetés, réalisés et consolidés par ses membres.

Si le programme de spécialisation en analyse des discours et de la communication publics est encore jeune, il repose sur une tradition de recherche conséquente et documentée qui nous autorise à donner quelques éléments relatifs aux types de corpus attendus.

L'analyse des discours et de la communication publics vise à rendre compte du rôle central du langage dans les interactions interpersonnelles médiatisées. Si elle s'inscrit dans le cadre général de l'analyse des pratiques verbales (discours et interactions), elle tire en effet sa spécificité de l'analyse des cas où ces pratiques font l'objet d'une médiatisation, c'est-à-dire où elles impliquent la mobilisation et la présence d'un public. L'analyse des discours et de la communication publics entend par conséquent donner non seulement une place conséquente aux enjeux spécifiques de ce type de pratiques, mais également aux ressources qui les rendent possibles et témoignent du travail médiatique[4].

[4] On prendra donc garde à ne pas prendre les données médiatiques pour argent comptant. Si ces données (exemplaires de journaux, émissions radiophoniques, débats télévisés, etc.) ont l'avantage d'être aisément accessibles au public et par conséquent au chercheur, cette accessibilité est la résultante d'un travail. Ces données sont fabriquées et dépendent par conséquent d'une pratique et d'enjeux spécifiques. La fabrication du produit médiatique a fait l'objet de plusieurs analyses en sciences du langage et de la communication. Prenons deux exemples : Broth (voir par exemple 2004 & 2008) a installé des caméras dans la régie d'une émission de plateau de manière à analyser la manière dont un débat télévisé est méthodiquement construit pour le

3.2. Objectif de recherche

Inscrite dans cette tradition disciplinaire, ma recherche doctorale entend *traiter de la parole politique en confrontation dans les débats publics pour considérer les dimensions et les enjeux identitaires que ces événements revêtent*. Cet intérêt se situe à l'intersection de deux soucis prégnants dans les sciences du langage et de la communication :

a) Contribuer à élaborer une approche socio-discursive de l'identité, en situant cette dernière à l'intersection du langagier (en tant que l'identité mobilise un *dire*) et du praxéologique (en tant que l'identité mobiliser un *faire*).

b) Contribuer à bâtir une approche socio-discursive des dispositifs interactionnels de confrontation politique (débat, discussion, polémique, etc.).

La formulation de ces objectifs oriente la constitution du corpus, l'analyste se mettant à la recherche de phénomènes spécifiques (qui ne constituent pas des données et donc encore moins un corpus). Ces phénomènes devront relever d'événements interactionnels publics (c'est-à-dire des événements où tous les participants sont co-présents ; il ne s'agit pas d'émissions télédiffusées) et orientés, plus ou moins contractuellement, vers la construction d'une confrontation verbale. Tel est l'air de famille, pour reprendre la célèbre expression de Wittgenstein, que devront avoir ces événements.

3.3. Quel critère de rassemblement de données en corpus ?

Dans un article publié en 2009, Vincent Capt, Raphaël Micheli et moi-même avions pour objectif de systématiser les critères mobilisés par le chercheur en analyse textuelle des discours lorsqu'il décide de rassembler – ce que nous avons décidé d'analyser comme un acte fort – deux phénomènes au moins dans un même ensemble de traitement, c'est-à-dire d'en faire un corpus. Nous avons identifié trois critères fondamentaux (le genre, l'auteur et le thème) et la possibilité de plusieurs articulations. Par exemple, la recherche de Micheli consacrée à l'analyse des débats parlementaires français sur l'abolition de la

télespectateur; Krieg-Planque (voir par exemple 2008a & 2008b) articule de son côté analyse linguistique et approche ethno-sociologique des pratiques journalistiques pour décrire l'élaboration et la mise en circulation de formules (par exemple « purification ethnique »).

peine de mort (2010) exploite les critères générique (débats parlementaires) et thématique (l'abolition de la peine de mort) pour garantir l'unification et la représentativité du corpus étudié. Dans le cas d'un tel corpus, c'est l'analyse des discours des différents auteurs (les députés) qui rend possible l'analyse contrastive (quels discours ont été tenus par les abolitionnistes et, réciproquement, par les anti-abolitionnistes).

Il ne s'agit pas d'entrer ici dans la complexité du problème. Il s'agit surtout d'encourager le chercheur à réfléchir aux principes qui le guident dans l'élaboration de ses corpus : « quel critère suis-je en train de mobiliser pour justifier la réunion en corpus de plusieurs phénomènes et, partant, quelle est ma motivation, que veux-je montrer ? » Dans mon cas, le critère qui guide ma recherche et par conséquent l'élaboration de mon corpus est essentiellement générique : mon but est de comprendre une pratique verbale (le débat public) et d'en saisir les enjeux.

3.4. Hétérogénéité constitutive des corpus en ADCP

Cet intérêt pour le genre du débat public ne doit pas faire oublier les impératifs de l'analyse des discours et de la communication publics (ADCP), et, parmi ceux-ci, l'attention portée non seulement aux pratiques médiatisées, mais également aux pratiques de médiatisation. La communication publique, qu'il s'agisse de débat public, de débat télévisuel, de conférence de presse, d'allocution télévisuelle, etc., implique la présence active d'une institution médiatrice dont le rôle ne se limite pas à organiser *in situ* l'événement et à s'assurer de sa « bonne » tenue, mais également à l'annoncer, c'est-à-dire à assurer, et ce très concrètement, son caractère public. Ainsi, un débat public n'est pas « déposé au milieu de nulle part », mais constitue en lui-même un aboutissement dont le corpus doit pouvoir fournir quelques traces. Parmi celles-ci, on trouvera donc, et ceci selon le genre médiatique considéré, des ressources aussi diverses que des affiches, des tracts, des annonces publicitaires, des courriers, des guides, des annonces informelles, etc.

Ce corpus complexe, projeté par la formulation d'objectifs visant la description des formes et la compréhension des enjeux d'une forme spécifique de communication publique, implique donc des sources hétérogènes, allant des événements publics eux-mêmes (les interactions orales de type « débat public » qu'il s'agit d'enregistrer) à leur environnement discursif (les divers documents

servant à les médiatiser et donc à garantir le caractère événementiel, ou « public », de ces rencontres).

4. Travail de terrain et corpus réalisé

Quel que soit le corpus projeté par la formulation d'un objectif de recherche, je fais l'hypothèse que le chercheur, du moment où il s'engage dans la phase de réalisation d'un corpus tel qu'il a été défini précédemment, s'engage inévitablement dans un véritable *travail de terrain*. Qu'il s'agisse d'archives disponibles dans une institution muséale, de textes ou documents audiovisuels disponibles sur Internet ou encore d'interactions orales authentiques, ces données ne diffèrent que du point de vue du support utilisé et par conséquent de leur accessibilité[5].

Dans tous les cas, le chercheur s'aventure et fait des choix. Il n'est donc pas inutile de rappeler les dimensions qu'implique un travail de terrain et les contraintes que celles-ci sont susceptibles de faire peser sur l'élaboration du corpus (4.1.). J'adopterai ensuite un mode plus narratif et listerai les différentes phases d'élaboration de mon corpus et les problèmes et choix qui se sont imposés et qui participent inéluctablement du corpus final (4.2-4.3.)[6]. Pour chaque étape, j'espère montrer que les paramètres et problèmes pertinents ne sont pas propres à mon travail, mais qu'ils sont ceux de l'analyse des discours et de la communication publics.

4.1. Aperçu des dimensions pertinentes d'un travail de terrain

a) Tout travail de terrain implique la *présence et la participation d'un observateur*. Ne pouvant compter sur la transparence de son observation, le chercheur ne doit jamais oublier qu'il constitue le premier principe organisateur de son corpus.

---

5 Qu'on me comprenne bien : je ne suggère pas de corrélation entre le type de support et la plus ou moins grande difficulté de l'accessibilité des données. Il n'est pas plus « facile » d'avoir accès à des archives muséales que d'obtenir l'autorisation et le matériel nécessaire pour enregistrer une conférence de presse. C'est bel et bien la nature de l'accessibilité qui diffère dans les deux cas.

6 L'architecture et les étiquettes des sous-sections de cette partie 4 sont largement inspirées des étapes et catégories proposées par le groupe ICOR (UMR 5191 ICAR) sur sa plate-forme de recherche CORINTE (icar.univ-lyon2.fr/projets/corinte). J'invite vivement le lecteur à consulter cette mine d'informations.

b) L'observateur-participant est un acteur social (un étudiant, un chercheur FNS, un professeur, etc.). Par conséquent, sa recherche s'inscrit dans un *contexte institutionnel, culturel et légal* qui contraint sa marge de manœuvre. Tout n'est pas possible, permis, bienvenu.

c) L'observateur doit gérer la manipulation nécessaire d'un *dispositif technologique* de captation et de fixation des données sur support. L'accessibilité et la disponibilité du matériel (carnet de notes, ordinateur avec accès internet, microphones, appareil photographique, caméra) ont une influence radicale sur les données recueillies et par conséquent sur le type d'analyse qu'elles rendent possibles.

d) L'utilisation d'un dispositif technologique exige un certain nombre de choix de la part de l'observateur-participant. Ceux-ci, dans le meilleur des cas, ne sont relatifs qu'aux objectifs de recherche qui guident alors seuls *le principe de pertinence* (*le* « *quoi capturer* »). Opération fondamentalement réductrice, l'enregistrement nécessite de la part du chercheur une prise en compte tant du déploiement spatial et temporel de l'interaction que des entités (individus, objets, documents) qu'elle mobilise.

4.2. Recueil des données

La phase de *recueil des données* débute avec la prise de contact avec le phénomène concerné et prend fin avec la fixation de données sur un support. Si elle nécessite des étapes similaires d'un corpus à un autre, cette phase peut largement différer quant à la longueur et la nature de ces étapes. Rappelons que les événements de type « débat public », par exemple, se caractérisent, entre autres choses, par leur dimension éphémère (vs enregistrée), localisée (vs médiatiquement diffusée) et agendée (vs spontanée).

Le premier enjeu du recueil d'un tel corpus réside par conséquent dans *la disponibilité du chercheur*. Celle-ci se situe à deux niveaux. Elle est spatiale d'abord, dessinant autour du chercheur un espace d'investigation (certains événements, aussi intéressants soient-ils, se situeront simplement « trop loin » ou même « ailleurs »). Elle est temporelle ensuite, dépendant de l'agenda du chercheur (certains événements, aussi intéressants soient-ils, tomberont simplement au mauvais moment). Dans le cas des débats publics, non seulement

la fréquence de ce type d'événements est faible (relativement à celle des conversations quotidiennes, des cours et séminaires universitaires, des réunions familiales, etc.), mais la position géographique et historique ainsi que la disponibilité du chercheur écartent donc de fait l'enregistrement de certains événements.

En second lieu, le recueil est contraint par la simple *découverte* de la tenue de tels événements. Le chercheur se trouve finalement dans la même situation que les futurs spectateurs : il prend connaissance (parfois trop tard) de la tenue d'un débat par voie d'affichage, de courrier ou de bouche-à-oreille. Ces documents (affiches et courriers) participent de la médiatisation des événements concernés. Il s'agit donc d'en conserver une trace ; ils font fondamentalement partie du corpus (cf. *supra*, 3.1. et 3.4.).

Figure 1 : L'affiche d'un débat.

Une fois qu'il a pris connaissance de la tenue d'un débat public et qu'il s'est assuré de sa disponibilité, le chercheur doit encore *prendre contact avec les organisateurs et les intervenants*. Cette prise de contact prend la forme d'une *demande d'autorisation*. Compte tenu du contexte légal dans lequel officie le chercheur, il n'est pas rare que des participants fassent appel à leur droit à l'image et se prononcent contre l'enregistrement. Une demande d'autorisation recouvre néanmoins bien plus qu'une simple procédure administrative. Il s'agit d'un lieu privilégié pour engager le dialogue entre Recherche et Cité. Le travail de terrain implique de la part du chercheur, et ceci d'un point de vue déontologique, un véritable effort de vulgarisation (la description des finalités vise l'obtention d'un consentement éclairé) et de prise en considération (l'écoute des préoccupations citoyennes permet d'identifier des objectifs de recherche potentiels). Les informateurs ont aussi droit à de l'information.

Dernière étape du recueil des données, et pas des moindres, *l'enregistrement* pose aux moins deux questions fondamentales : celle du dispositif et celle de la pertinence. Opération à la fois matérielle (en tant qu'elle mobilise un dispositif technologique) et technique (en tant qu'elle nécessite des manipulations et par conséquent une pratique), l'enregistrement exige de la part du chercheur la prise en compte de l'ensemble des dimensions de l'interaction verbale (son

déploiement dans l'espace et le temps ainsi que les cadres participatifs qu'elle mobilise). Le choix du type d'enregistrement (prise de notes, enregistrement sonore ou audiovisuel) dépend essentiellement des moyens à disposition et des autorisations accordées par les participants et par le contexte institutionnel et culturel. Il est relatif au degré d'intrusion que les participants et/ou le chercheur sont prêts à accepter. L'enregistrement est fondamentalement articulé à la problématique de la pertinence.

Du fait qu'elle exige le choix d'un point focal et par conséquent d'un cadrage, l'opération ajoute à la première réduction, celle imputable au choix d'un événement à enregistrer (parmi tous les autres possibles), une seconde, liée à la nécessité d'organiser le champ *versus* hors-champ. Avec l'enregistrement, le chercheur se trouve de facto confronté à un jeu de double contrainte et à la nécessité de compromis. D'une part, du fait qu'il entend décrire des événements dans leur spécificité (par exemple des débats publics), l'analyste de la communication publique est tenté d'en privilégier une saisie radicalement englobante et externe. D'autre part, sensibles à la pertinence vers laquelle s'orientent les participants aux interactions, les analyses minutieuses que le même analyste entend mener nécessitent tant la compréhension et l'attribution de ce qui se dit et de ce qui se passe, que la possibilité d'endosser la perspective

Figure 4 : Captures et points de vue

des participants, c'est-à-dire la manière dont les agents sociaux impliqués perçoivent l'interaction et y participent. Le chercheur est donc tiraillé entre la description externe des situations de communication dans lesquelles sont engagés les agents et la description, interne, de leurs points de vue. D'un côté, les inconvénients d'une capture large résident dans la non-représentativité des données (cette perspective n'est celle de personne, si ce n'est celle du chercheur) et dans un risque d'indiscernabilité (qu'il s'agisse d'identifier une

voix ou de reconnaître un geste). De l'autre, la capture étroite risque d'altérer plus fondamentalement la dynamique communicative que ne le ferait la saisie englobante (le dispositif est plus visible, plus présent) et de passer sous silence ou dans un angle mort des phénomènes pertinents qui échapperont ainsi à l'enregistrement et de surcroît à l'analyse. Dans mon cas, l'opportunité de disposer de deux caméras (couplées à deux dictaphones munis de microphones performants) a permis de jouer avec la complémentarité de prises de vues différentes. Généralement, une caméra était réservée à la capture de la « scène » (intervenant(s) invité(s) et organisateur) à partir du point de vue du public tandis que l'autre, disposée à une extrémité de la scène, permettait d'enregistrer le comportement du public du point de vue des intervenants invités[7]. On le conçoit aisément, l'enregistrement dépend essentiellement des autorisations obtenues, mais aussi du matériel à disposition et par conséquent de l'inscription du chercheur dans des réseaux de chercheurs, dans une institution mettant du matériel à disposition de son personnel, etc.

La fixation sur support d'une certaine quantité de données est inhérente à tout travail sur corpus. Qu'il s'agisse de données manuscrites, imprimées, numériques ou même multimédias, le chercheur est toujours concerné par une réflexion sur les relations qu'entretiennent le phénomène qu'il désire se donner comme objet d'analyse et la fixation de celui-ci sur un support et dans un format adaptés à la tâche. Le stockage de données dans un dispositif d'enregistrement ne constitue néanmoins qu'un préalable à la phase de confection.

4.3. Confection des données

On entend par confection des données les différentes étapes visant à préparer la postérité des données enregistrées en vue d'une gamme d'applications (enseignement, recherche, publications, vulgarisation, etc.).

---

[7] L'enregistrement du premier événement fait exception à la norme. Dans ce cas, les deux caméras étaient dirigées sur les intervenants invités, laissant le public hors champ. Deux éléments me semblent expliquer ce choix discutable : premièrement, les intervenants étant au nombre de dix, une seule caméra (l'autre étant alors réservée au public) n'aurait pas permis une prise de vue optimale de leur comportement discursif et interactionnel (soit il aurait fallu zoomer sur la personne s'exprimant au risque de mettre hors champ les autres invités ; soit il aurait fallu choisir un cadrage très large et englobant au risque de ne pouvoir discerner grand chose). Deuxièmement, et cela est plus intéressant pour mon propos, la prise de vue choisie m'a fait prendre conscience d'un préconstruit tenace : le public n'est pas important ou du moins n'est pas si important. Cette « erreur » ainsi que l'expérience générale de l'enregistrement m'ont beaucoup apporté. Elles conduisent à considérer les données médiatiques (les images produites par les médias) en tant que véritables pratiques de cadrage et de sélection de la pertinence.

Première étape de la confection des données : *la numérisation*. Aujourd'hui essentielle, la numérisation vise à *assurer à la fois la pérennité des données enregistrées, leur interrogeabilité et leur compatibilité* avec divers logiciels d'aide à l'analyse[8]. Opération matérielle, la numérisation implique le choix d'un format. On préconise généralement de choisir les formats les plus courants et les plus compatibles (de manière assurer la pérennité et l'interopérabilité des données), les moins propriétaires (de manière à ne pas être tributaire, financièrement et institutionnellement parlant, d'entreprises dont la pérennité n'est pas assurée) et les plus performants (de manière à trouver le meilleur compromis entre taille et qualité du fichier cible).

Une fois la pérennité des données assurée, la question de *leur anonymisation* est heureusement rapidement réglée. Du fait du caractère public des débats auxquels ils participent et qui s'opposent aux interactions ayant un caractère privé (conversation quotidienne, consultation médicale, réunions professionnelles, etc.), les intervenants acceptent contractuellement de parler en leur nom ; ils engagent une identité publique. C'est là une spécificité des contextes relevant des discours et de la communication publics.

L'opération de *transcription* est à la fois nécessaire, déterminante et délicate. Enjeu de confrontation théorique, elle fait l'objet d'une abondante bibliographie. Pour des raisons de place, je me contente de rappeler quelques éléments fondamentaux[9]. On entend par transcription la production d'un texte discret – le continu de l'expérience est segmenté en unités compositionnelles – visant une certaine transparence avec les données originales (orales ou même multimodales) [10] . Fondamentalement réductrice, la transcription est un appauvrissement permettant de réduire la complexité initiale en vue d'une

8 Si la plupart des dispositifs d'enregistrement sont aujourd'hui numériques (ordinateur, appareil photographique, caméra) et ne nécessitent par conséquent pas d'être couplés à des dispositifs de numérisation, ceux-ci peuvent se révéler nécessaires, et ce pour des raisons diverses (coûts des dispositifs numériques, formats des documents archivés, problèmes d'accessibilité aux fichiers numériques sources). Quelques exemples : reproduction à l'ordinateur d'une correspondance manuscrite ; utilisation d'un scanner et d'un logiciel OCR (reconnaissance de caractères) pour numériser une lettre commerciale, une brochure électorale, un exemplaire de journal ou encore une affiche publicitaire ; numérisation de matériel audio ou audiovisuel.

9 A la fin de la présente contribution, la section « bibliographie sélective et références » regroupe bon nombre de références consacrées aux problèmes et défis que posent la transcription, et ce dans des domaines variés des sciences du langage et plus généralement des sciences humaines et sociales. On se réfèrera par exemple aux travaux de Ochs 1979 ; Blanche-Benveniste et Jeanjean 1987 ; Psathas et Anderson 1990 ; Edwards et Lampert 1993 ; Ashmore et Reed 2000 ; Bucholtz 2000 ; Mondada 2000 & 2002 ; Ten Have 2002 ; Jefferson 2004 ; Dister et Simon 2007). En outre, le Département de Linguistique de l'Université de Californie consacre une abondante bibliographie au sujet (http://www.linguistics.ucsb.edu/projects/transcription/bibliography).

10 Contrairement à une idée relativement répandue, la transcription, tout comme le corpus, ne constitue pas la base inflexible de l'analyse. Si elle est nécessaire à l'analyse, elle en constitue également un résultat.

gamme d'applications (décrire, présenter, analyser, discuter un phénomène). Le « quoi transcrire » et le « comment transcrire » (les spécialistes parlent de niveaux de *granularité*) dépendent par conséquent du « quoi enregistrer » et donc des objectifs et traditions de recherche. Ces dernières impliquant un dialogue entre chercheurs, la nécessité d'une certaine homogénéité a abouti à une sédimentation des pratiques, sous forme de conventions de transcription[11].

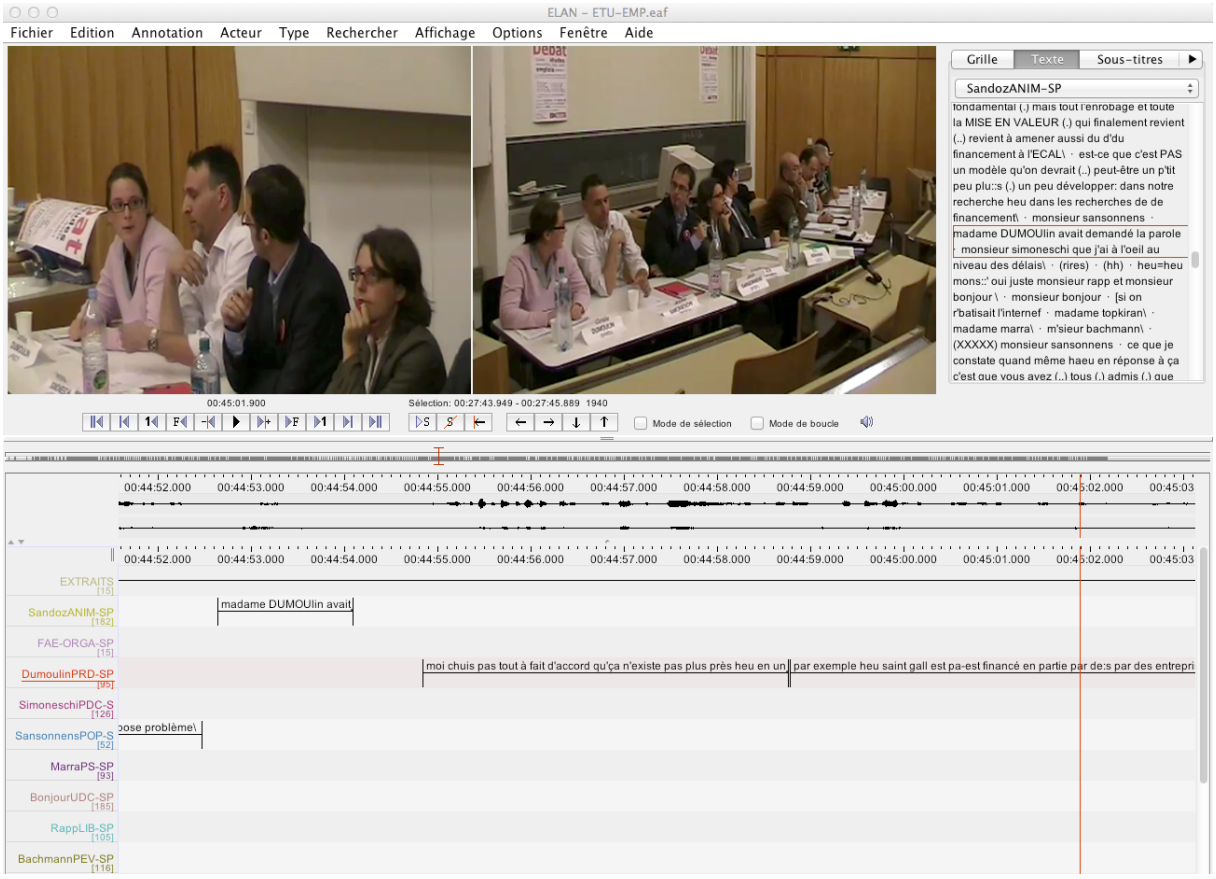
Arrivé à ce stade, le chercheur possède d'une part une certaine quantité de données audio ou audiovisuelles et d'autre part des données textuelles (les transcriptions)[12]. Outre le risque que le chercheur délaisse les premières au profit des secondes, un problème d'accessibilité se pose : il n'y a pas à ce stade de coordination automatique dans la consultation des données textuelles et audio(visuelles) : une fois un passage intéressant relevé dans la transcription, il faut fastidieusement rechercher l'endroit correspondant dans les données originales et jongler d'une source à l'autre. Pour pallier ces problèmes, il existe depuis quelques années des logiciels rendant possible *l'alignement* – c'est-à-dire la synchronisation – de l'ensemble des composantes. Que ce soit sous forme de partition de musique, de sous-titres ou encore de texte classique souligné par surbrillance, les logiciels en question permettent d'afficher la transcription correspondant au moment du fichier source que le chercheur consulte. Les diverses dimensions du corpus sont ainsi rendues solidaires[13].

11 Pour ma part, j'ai opté pour les conventions de transcription du groupe ICOR (disponibles à l'adresse suivante : http://icar.univ-lyon2.fr/projets/corinte/bandeau_droit/convention_icor.htm). Les conventions ICOR possèdent un triple avantage : (*i*) elles s'inspirent des conventions classiquement usitées en analyse conversationnelle (Jefferson 2004), et plus généralement en analyse des interactions verbales ; elles sont par conséquent aisément et rapidement intelligibles par un lecteur familier des transcriptions ; (*ii*) elles sont utilisées pour des objectifs et des analyses analogues au présent travail, elles ont donc déjà été éprouvées ; (*iii*) elles proposent des conventions de transcriptions de la mimo-gestualité.

12 Dans mon cas, il y a en fait trois types de données : les données audio tirées des microphones, les données vidéos tirées des caméras et les transcriptions. En outre, n'oublions pas les ressources de médiatisation (annonces diverses de l'événement). Si elles font fondamentalement partie du corpus, ces données ne sont pas concernées par les étapes dont il est question ici. Elles sont par contre ré-exploitées lorsque la constitution du corpus laisse place à sa présentation et à son analyse.

13 L'analyse de la prosodie (pour l'articulation texte – audio) ou de la mimo-gestualité et de la proxémie (pour l'articulation texte – vidéo) est alors grandement facilitée.

Figure 3: Le logiciel d'alignement Elan



La vie d'un corpus est scandée par des moments où sa réalisation est interrompue et son état stabilisé. Ces moments (travaux pratiques, examens, publications diverses, défense de mémoire ou soutenance de thèse) nécessitent de la part du chercheur une réflexion sur les modes de présentation des données et sur les manières d'expliciter et justifier la construction du corpus. Ce sont des moments où la réalisation du corpus laisse place à sa systématisation en vue de l'analyse des données et la présentation des résultats, où le chercheur opère un retour réflexif sur la dialectique entre projection et réalisation qui l'a guidé lors de l'élaboration du corpus. Fonction de cette dialectique, la phase de consolidation est donc propre à chaque corpus. Elle sort de la question stricte de l'élaboration des corpus et par conséquent du cadre de la présente contribution.

5. Conclusion

Cette contribution avait pour objectif la présentation des enjeux que revêt l'élaboration des corpus en analyse des discours et de la communication publics. Pour leur donner une certaine consistance, j'ai pris l'exemple de mon corpus de recherche et en ai balisé les étapes de constitution. Adoptant comme bien d'autres une perspective résolument praxéologique, j'ai considéré que corpus et données étaient fonction d'un objectif de recherche et d'une pratique de terrain :

le corpus est un construit en vue d'une gamme d'applications. Si ce constat est aujourd'hui relativement banal, il s'agissait surtout d'en tirer les conséquences pour l'analyse des discours et de la communication publics. Outre le caractère multimodal que revêt fréquemment ce type de phénomènes et qui nécessite une méthodologie et un dispositif technique appropriés, une des conséquences majeures consiste en la nécessité, impliquée par le caractère public des événements de parole concernés, de conserver une trace du travail de publicisation et de médiatisation. Du fait de leur hétérogénéité, les corpus de communication publique sont par définition complexes. En tirer toutes les conséquences méthodologiques et théoriques reste une tâche cruciale.

Références bibliographiques

ASHMORE, Malcolm et Darren REED (2000), « Innocence and Nostalgia in Conversation Analysis: The Dynamic Relations of Tape and Transcript », *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research,* vol. 1, n°3, http://www.qualitative-research.net/index.php/fqs/article/view/1020/2200.

BAUDE, Olivier, Claire BLANCHE-BENVENISTE, Marie-France CALAS *et al.* (2006), Corpus oraux, guide des bonnes pratiques 2006, HAL - CCSD.

BERTHELOT, Jean-Michel (2000), *Sociologie : épistémologie d'une discipline*, Bruxelles, De Boeck Université.

BIBER, Douglas, Susan CONRAD et Randi REPPEN (1998), *Corpus Linguistics, Investigating Language Structure and Use*, Cambridge, CUP.

BILGER, Mireille (Ed.) (2000), *Corpus. Méthodologie et applications linguistiques*, Paris, Champion.

BLANCHE-BENVENISTE, Claire et Colette JEANJEAN (1987), *Le français parlé. Transcription et édition*, Paris, Didier Erudition.

BOURDIEU, Pierre, Jean-Claude CHAMBOREDON et Jean-Claude PASSERON (2005), *Le métier de sociologue : Préalables épistémologiques*, Paris, EHESS ; Mouton de Gruyter.

BROTH, Mathias (2004), « La réalisation des images télévisées comme accomplissement social », dans Jeanne STRUNCK, Inger LASSEN et Torben VESTERGAARD (Eds.), *Genres et rhétorique des discours médiatiques*, Aalborg, Aalborg University Press, p. 107-123.

BROTH, Mathias (2008), « The studio interaction as a contextual resource for TV-production », *Journal of Pragmatics*, n°40, p. 904-926.

BUCHOLTZ, Mary (2000), « The politics of transcription », *Journal of Pragmatics,* vol. 32, n°10, p. 1439-1465.

CAPT, Vincent, Jérôme JACQUIN et Raphaël MICHELI (2009), « Les sphères de contextualisation : réflexion méthodologique sur les passages de texte à texte(s) et la constitution des corpus », Corpus, n°8, p. 129-147.

DISTER, Anne et Anne-Catherine SIMON (2007), « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica*, vol. 1, n°1, p. 54-79.

DURANTI, Alessandro et Charles GOODWIN (Eds.) (1992), *Rethinking context language as an interactive phenomenon*, Cambridge, Cambridge University Press.

EDWARDS, Jane et Martin LAMPERT (Eds.) (1993), *Talking data: Transcription and coding in discourse research*, Hillsdale, NJ, Lawrence Erlbaum.

EMERSON, Robert (1997), « Le travail de terrain après Hughes : continuités et changements », *Sociétés contemporaines*, n°27, p. 39-48.

GOFFMAN, Erving (1989), « On Field Work », *Journal of Contemporary Ethnography*, n°18, p. 123-132.

HABERT, Benoît, Adeline NAZARENKO et André SALEM (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

JACQUES, Marie-Paule (2005), « Pourquoi une linguistique de corpus ? », dans Geoffrey WILLIAMS (Ed.), *La linguistique de corpus*, Paris, PUR, p. 21-30.

KENNEDY, Graeme (1998), *An Introduction to Corpus Linguistics*, London; New York, Longman.

KRIEG-PLANQUE, Alice (2008a), *« Purification ethnique ». Une formule et son histoire*, Paris, CNRS Editions.

KRIEG-PLANQUE, Alice (2008b), « La notion d'"observables en discours". Jusqu'où aller avec les sciences du langage dans l'étude des pratiques d'écriture journalistique ? », dans Marcel BURGER (Ed.), *L'analyse linguistique des discours médiatiques*, Québec, Nota Bene, p. 59-81.

JEFFERSON, Gail (2004), « Glossary of transcript symbols with an Introduction », in Gene H. LERNER (Ed.), *Conversation Analysis: Studies from the first generation*, Philadelphia, John Benjamins, p. 13-23.

MAYAFFRE, Damon (2002), « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus,* n°1, p. 51-69.

MAYAFFRE, Damon (2005), « Rôle et place des corpus en linguistique », *Actes des Journées d'Etude TOUlousaines JETOU*, p. 5-17.

MELLET, Sylvie (2002), « Corpus et recherches linguistiques », *Corpus,* n°1, p. 5-12.

MONDADA, Lorenza (1998), « Technologies et interactions dans la fabrication du terrain du linguiste », *Cahiers de l'ILSL*, n°10, p. 39-68.

MONDADA, Lorenza (2000), « Les effets théoriques des pratiques de transcription », *LINX*, n°42, p. 131-150.

MONDADA, Lorenza (2002), « Pratiques de transcription et effets de catégorisation », *Cahiers de Praxématique*, n°39, p. 45-75.

MONDADA, Lorenza (2005), « L'analyse de corpus en linguistique interactionnelle : de l'étude de cas singuliers à l'étude de collections », dans Anne CONDAMINE (Ed.), *Sémantique et corpus*, Paris, Hermès, p. 76-108.

MONDADA, Lorenza (2006a), « Video Recording as the Preservation of Fundamental Features for Analysis », dans Hubert KNOBLAUCH, Jürgen RAAB, Hans-Georg SOEFFNER *et al.* (Eds.), *Video Analysis*, Berne, Peter Lang, p. 51-68.

MONDADA, Lorenza (2006b), « La pertinenza del dettaglio : registrazione e trascrizione di dati video per la linguistica interazionale », in Yvette BÜRKI et Elwys DE STEFANI (Eds.), *Trascrivere la lingua. Dalla filologia all'analisi conversazionale*, Bern, Peter Lang, p. 313-344.

MONDADA, Lorenza et Christian PLANTIN (2006), *Site internet CORINTE (Corpus d'Interactions) liée au projet CLAPI (Corpus de Langue Parlée en Interaction) du laboratoire ICAR (Interactions, Corpus, Apprentissages, Représentations)*, consultable sur http://icar.univ-lyon2.fr/projets/corinte/.

NORRIS, Sigrid (2004), *Analyzing multimodal interaction : a methodological framework*, New York, Routledge.

OCHS, Elinor (1979), « Transcription as theory », in Elinor OCHS et Bambi SCHIEFFELIN (Eds.), *Developmental pragmatics*, New York; San Francisco; London, Academic Press, p. 43-72.

PSATHAS, George et Tim ANDERSON (1990), « The 'practices' of transcription in conversation analysis », *Semiotica,* vol. 1-2, n°78, p. 75-100.

RASTIER, François (1998), « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, n°128, p. 97-111.

RASTIER, François (2004), « Enjeux épistémologiques de la linguistique de corpus », revue en ligne *Texto!,* http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.

TEN HAVE, Paul (2002), « Reflections on transcription », *Cahiers de praxématique*, n°39, p. 21-43