Du Français Fondamental aux ESLO

Lotfi Abouda, Olivier Baude

Du Français Fondamental aux ESLO

ABOUDA, Lotfi & BAUDE, Olivier
Université d'Orléans

L'Enquête Socio-Linguistique à Orléans (ESLO), réalisée en 1968-71, marque une rupture méthodologique et théorique avec le Français Fondamental sur différents points qui sont représentatifs des différents usages des corpus oraux en linguistique. Ainsi, la définition des enjeux de l'exploitation scientifique - numérisation, transcription, annotation, diffusion, analyses - du corpus ESLO conçue comme étape liminaire à la réalisation d'une nouvelle enquête variationniste en cours d'élaboration (ESLO2), permet d'interroger l'évolution des modèles et des méthodes de constitution et d'exploitation des corpus oraux destinés à des finalités linguistiques et didactiques.

Mots-clés : corpus oraux, linguistique variationniste, numérisation, transcription

The "*Enquête Socio-Linguistique à Orléans*" (ESLO), conducted from 1968 to 1971, marks a methodological and theoretical break with the "*Français Fondamental*" on the different levels corresponding to the different uses made of oral corpora. Hence, re-defining the scientific stakes - digitalization, transcription, annotation, distribution, analyses - involved in the analysis of the ESLO corpus (in fact the first stage of a new project of variationist data collecting, ESLO 2) will allow us to question the models and methods employed in constituting and exploiting oral corpora intended for linguistic and didactic purposes.

Keywords : Spoken corpora, variationist linguistics, digitalization, transcription

0. Introduction

Une dizaine d'années après la réalisation du Français Fondamental (FF), une équipe d'universitaires britanniques a entrepris la constitution d'un corpus de français oral, à visée à la fois linguistique et didactique : l'Enquête Socio-Linguistique à Orléans (désormais ESLO).

L'ESLO marquera une rupture avec plusieurs décisions ayant guidé la réalisation du Français Fondamental : souci d'une identification sociologique raisonnée en termes de CSP, conservation des enregistrements, adaptation du corpus à des interrogations multiples, préservation de la cohérence discursive, réflexivité de l'enquête, observation de l'interaction et des conduites linguistiques…

L'objet de cet article sera de présenter les principales étapes du travail de reconstruction d'ESLO1, actuellement en cours au sein du Centre Orléanais de Recherche en Anthropologie et Linguistique (CORAL), reconstruction vue comme une étape liminaire à la réalisation d'une nouvelle enquête variationniste à Orléans, ESLO2.

En comparant ESLO1 reconstruit au FF, nous pouvons mesurer tout le chemin parcouru en cinquante ans, ce qui ne se résume pas à une simple évolution technique, même si celle-ci y a joué un rôle décisif.

1. Du français fondamental à l'ESLO

1.1. Situations et contextes historiques

Souvent comparé, voire vu comme une réponse, au *Basic English*, le « Français Fondamental » (FF) a été créé au début des années 1950, pour promouvoir le français dans des colonies qui allaient bientôt devenir indépendantes.

La décision de recueillir du français oral, située dans son contexte épistémologique – très largement dominé par l'écrit –, peut s'expliquer par l'évolution de l'enseignement du français langue étrangère (FLE). En effet, l'efficacité de l'apprentissage du FLE, devenu une véritable technique évaluable en termes de coût, « suppose, ainsi que l'écrivent Bergounioux et al. (1992, p. 75-76), une méthodologie particulière, fondée sur des relevés quantifiés moins dépendants de la littérature classique ».

En 1966, un groupe de linguistes britanniques se réunit pour établir un bilan de l'enseignement du FLE en Angleterre. Si, dans ce pays, le rôle de l'oral dans l'enseignement des langues étrangères était reconnu depuis longtemps, il manquait pour le français un ensemble cohérent, systématique et actualisé de matériaux pédagogiques.

L'inexistence, y compris en France, de matériaux exploitables poussa ces chercheurs à entreprendre la constitution d'un nouveau corpus. L'entreprise, qui dura plus de cinq ans entre la conception et la réalisation, donna naissance à l'un des corpus les plus vastes de français oral : 350 bandes magnétiques représentant quelques 317 heures d'enregistrements, ce qui correspond à quelques ± 4 500 000 mots.

Mais, ainsi qu'on le verra, la taille de ce corpus ne constituera pas son seul intérêt.

1.2. A la recherche de données spontanées et situées

Si les objectifs déclarés de l'ESLO sont multiples, sa principale raison d'être reste tout de même didactique – l'enseignement du FLE –, objectif qu'elle partage avec le FF. Mais au-delà de toutes les différences didactiques, nombreuses, qui opposent les deux projets, la principale opposition concerne la nature elle-même des matériaux recueillis.

En effet le FF, s'il a bien intégré la langue orale, le fait dans le but de configurer un état moyen du français. L'oral a été ici homogénéisé, figé, fragmenté, ramené à des moyennes, avant de totalement disparaître.

Sur ce point, les initiateurs d'ESLO ont pris des décisions diamétralement opposées, en choisissant de représenter la variété des usages.

Cette variété des usages, revendiquée – dans le champ alors balbutiant de la sociolinguistique – et concrètement intégrée dans ESLO comme autant de variables, concernait aussi bien les différences générationnelles et communautaires, que les différences sociales, sans négliger les différences liées aux conditions de production du discours (Blanc & Biggs 1971, p. 16).

1.2.1. Quand le « Portrait sonore d'une ville » cache une « communauté d'auditeurs »

Pour faire accepter l'enquête auprès de la population locale, les enquêteurs ont dû mettre en avant le concept de « portrait sonore d'une ville ».

En réalité, le choix d'Orléans fut, sur le plan géographique, un non choix. Pour écarter des variables incontrôlables, il fallait une ville, de la taille d'Orléans, dont on pouvait reconstruire la dynamique des formes linguistiques simultanément présentes, une cité « assez vaste pour que la variation y soit accusée et perpétuée à travers des réseaux linguistiques d'échanges autonomes, et assez restreinte pour que n'importe quel membre de cette communauté linguistique ait dû interférer dans les circuits de communication des autres groupes. » (Bergounioux et al. 1992, p. 79). Cela écarte l'hypothèse qui ferait du choix d'Orléans celui d'une forme standardisée du français, représenté par une variété régionale non marquée. Et de fait, on rencontre parmi les locuteurs enregistrés un nombre non-négligeable de témoins qu'une enquête dialectologique aurait écartés[1]. L'accent de certains de ces témoins était vu par les enquêteurs non pas comme une imperfection, mais comme un intérêt supplémentaire de l'enregistrement.

Ces faits appuient fortement l'hypothèse formulée par Bergounioux et al. (1992) quand ils écrivent (op. cit.) : « C'est une communauté d'auditeurs qui est construite, autant qu'une communauté de locuteurs, à notre connaissance pour la première fois en France. » En bref, il s'agissait pour les enquêteurs de saisir les variétés du français qu'on pouvait entendre à un moment donné dans un lieu donné, non marqué géographiquement.

A l'intérieur de ce cadre, l'objectif des enquêteurs n'était pas de rechercher « cet individu mythique, l'Orléanais moyen » (Blanc & Biggs 1971, p. 23), mais d'appréhender dans toute son hétérogénéité une communauté de locuteurs-auditeurs.

De telles exigences nécessitaient le recours à une véritable enquête linguistique dont la méthodologie était à l'époque encore à construire.

[1] Trois Pied-Noirs d'Algérie, deux Aquitains, deux Bretons, un Lorrain et dix Parisiens.

1.2.2. Panel et échantillonnage

Une fois la ville choisie, il fallait affronter le délicat problème de la détermination de l'échantillon de population sur lequel porterait l'enquête.

Les initiateurs d'ESLO, retenant la méthode de l'échantillonnage au hasard, ont fait appel à l'INSEE pour réaliser une sélection aléatoire de 600 témoins, répartis, en plus du sexe, en 3 tranches d'âges (18/30, 31/50, 51 et +) et 6 catégories socioprofessionnelles. Cet échantillon ne se voulait pas représentatif mais diversifié et offrant un nombre égal suffisant de témoins pour une *étude linguistique*.

Si, sur le principe, la démarche se distinguait nettement de celle du FF, les résultats concrets n'étaient pas à la hauteur des espérances, puisque le taux de refus, prévu autour de 50%, a été très largement dépassé (au final, ils n'ont obtenu que 147 entretiens, soit moins de 25%). Ce rendement, faible, ne pouvait que nuire à l'équilibre de l'échantillonnage (par exemple, il y a eu peu d'ouvriers, dont la plupart étaient de surcroît rompus à la prise de parole publique).

Mais la technique d'échantillonnage, tout en ouvrant la porte à une enquête rigoureuse, ne pouvait qu'endommager les méthodes d'enquêtes.

Ainsi, dans un même souci de comparabilité, l'équipe privilégiera la situation de l'entretien en face à face, situation certes très formelle, mais qui avait l'avantage d'être pour eux contrôlable. « *Les mêmes questions sont posées par les mêmes personnes dans les mêmes conditions* », écriront Blanc & Biggs (op.cit., p. 17).

1.2.3. A la recherche de formes spontanées

Cette volonté de rigueur méthodologique dans le contrôle de la situation et dans la restriction des variables s'est confrontée à l'objectif de recueillir du français spontané dans un bel exemple de ce que la sociolinguistique nommera par la suite "le paradoxe de l'observateur"[2].

Les initiateurs d'ESLO chercheront alors à compléter le corpus d'entretiens par le recours à des situations plus naturelles, qu'ils paraissaient toutefois ne pas totalement maîtriser, à la fois sur le plan de la méthodologie de l'enquête et sur celui de la qualité technique de l'enregistrement.

Ces emplois plus spontanés représentent in fine 2/3 des enregistrements regroupés en cinq catégories:

1. des reprises de contacts informelles (15 reprises de contacts dans des situations variées : discussion entre amis,…)
2. enregistrements au hasard en micro caché (rue, magasins, etc.)
3. interviews de personnalités du monde syndical, politique, universitaire et de l'administration d'Orléans (maire, évêque, …)
4. tables rondes, conférences-débats (sur des thèmes variés : condition de la femme, promotion sociale,…)

[2] Labov 1973.

5. entretiens au Centre Médico Psychopédagogique d'Orléans (entretien entre une assistance sociale et des parents).

1.2.4. De Bernstein à Bourdieu

A ce tâtonnement (technique et théorique) dans la recherche du recueil de français spontané en interaction, s'est ajouté celui d'une sociolinguistique applicable à l'enquête.

L'entretien en face-à-face a été élaboré sur la base de 3 questionnaires : le premier (ouvert) devait permettre de recueillir les positions du témoin sur son expérience personnelle et divers types de discours déterminés. Le second, semi-fermé, intitulé "questionnaire socio-linguistique" et confié à un élève de Pierre Bourdieu – Bernard Vernier –, s'il porte encore les traces des théories de Bernstein sur la langue[3], enferme un recueil des représentations du témoin sur la norme linguistique et culturelle. La sociologie naissante de Bourdieu se rencontre également dans le troisième questionnaire, fermé, qui porte, parallèlement à l'état civil, sur les pratiques déclarées des habitudes culturelles.

Cette importance donnée au capital culturel s'est également concrétisée dans l'élaboration d'une nouvelle grille, l'échelle AM (de son concepteur Alix Mullinaux, qui comprend cinq agrégats (notés de A à E). Complémentaire à celle de l'INSEE, cette grille tente de rendre compte, parallèlement aux critères habituels, des pratiques et références culturelles ainsi que de la mobilité géographique potentielle des témoins.

Ce n'est pas le moindre des intérêts du projet ESLO que de porter les traces de la fin d'une sociolinguistique militante (et naïve) qui se bornait à corréler variations sociales et hiérarchisation de la compétence linguistique, et les prémices d'une nouvelle sociologie qui donnera toute sa place à la distinction apportée par le capital culturel.

1.3. Disponibilité des données

1.3.1. Français fondamental et données volatiles.

Le français fondamental était fondé dès son origine sur le constat de la pauvreté des enregistrements du français parlé disponibles, que ce soit dans les fonds des institutions de conservation du patrimoine que dans ceux des archives scientifiques[4].

Mais, lorsqu'ils ont entrepris de recueillir leurs propres données, ils n'ont pas hésité à effacer les enregistrements audio au fur et à mesure qu'ils les ont transcrits.

Comment expliquer ce choix, étonnant, qui revenait de fait à dénier à ces enregistrements toute reconnaissance comme objet scientifique et patrimonial ?

[3] « L'origine sociale et le niveau d'éducation du témoin concourent avec ses activités socioprofessionnelles à déterminer sa compétence linguistique. Selon cette hypothèse plusieurs types d'attitudes seraient en corrélation avec les niveaux socio-culturels ». Blanc & Biggs (op. cit., p. 18)

[4] « Entre le peu d'intérêt du Musée de la Parole pour des enregistrements de français, la dominante folklorique des documents conservés au Musée des arts et traditions populaires (2 documents exploitables) et le caractère gourmé des interventions en dépôt aux Archives de la radio (6 documents exploitables), la moisson fut maigre ». Bergounioux et al. (1992, p. 76).

La première raison est purement technologique. Les enregistrements du Français Fondamental ont été réalisés sur des disques de papier magnétique particulièrement fragiles[5], et considérés comme coûteux.

Mais la vraie raison est sans doute ailleurs : l'équipe du FF pensait que la transcription, qu'elle voulait rapide, était un moyen de conserver ce corpus, sans que ne soit évoquée la question de la réutilisation des données primaires, par d'autres ou par l'équipe elle-même, ne serait-ce que par soucis de contrôle.

1.3.2. L'ESLO : des données résolument disponibles

Sur ce point, aussi, le projet ESLO exposera des choix diamétralement opposés. Il s'agissait bel et bien de constituer un corpus *disponible* pour de multiples travaux, aussi bien en linguistique qu'en didactique du FLE, dans une démarche inhabituelle en France.

La conservation et la possibilité de réutilisation des matériaux recueillis étaient dès le départ considérées par les initiateurs d'ESLO comme deux objectifs fondamentaux. Ce choix s'est concrétisé de différentes manières.

1. D'abord par le **catalogage** et l'**indexation** : l'équipe de d'ESLO a publié en 1974 un catalogue descriptif et analytique[6] qui répertoriait les enregistrements avec : résumé du contenu, indexation des questions, organisation du questionnaire, catégorisation sociologique précise des locuteurs et description de la situation d'enquête.
2. **La conservation des données primaires (enregistrements et documents d'enquête).**
3. **Les transcriptions.** Bien qu'une transcription intégrale fût difficilement envisageable pour un corpus estimé à plus de 4 millions de mots, l'équipe a entrepris immédiatement la transcription d'extraits qui se voulaient représentatifs et qui recouvraient toutes les catégories des témoins (INSEE et AM).
4. **La diffusion du corpus**

   Outre l'annonce systématique dans les articles de la disponibilité du corpus, le catalogue précise dès la page 4:

> "*Les transcriptions et enregistrements sont disponibles à tout chercheur intéressé, contre remboursement des frais de matériaux et de copiage; (…) Des listes de transcriptions et enregistrements sont disponibles à ceux qui s'adressent à nous.*" (Lonergan,et al.,1974, p. 4).

[5] Cependant cette limitation de la technologie est presque présentée comme un avantage : « L'appareil Recordon que nous avons choisi présentait pour notre travail des avantages certains. Son poids était peu élevé : un peu plus de 6 kg. […]. L'interruption dans les conversations que nous imposait le changement des disques toutes les six minutes, au terme de leur durée maxima d'enregistrement, ne présentait pas non plus d'inconvénient. Nous ne nous souciions pas non plus de la conservation des disques. Nous profitions largement des possibilités qu'offrent les disques en papier d'être effacés et de servir ainsi à plusieurs enregistrements successifs. Il aurait été beaucoup trop coûteux de conserver tous les enregistrements comme beaucoup de bons esprits nous le suggéraient » Gougenheim et al. (1964, p. 63).

[6] Il s'agit d'un important volume de 218 pages, reproduit en 1993 par l'université d'Orléans.

1.4. Un corpus peu exploité

Les initiateurs d'ESLO auront donc tout fait pour rendre leur corpus disponible, et exploitable. Cela n'a pas empêché que son exploitation didactique fût faible[7], et son exploitation scientifique pendant longtemps quasi inexistante, en tout cas en France.

Le diagnostic de cet échec – eu égard à l'ambition initiale du projet – fait apparaître plusieurs causes.

D'abord, ESLO constitue un corpus particulièrement encombrant et lourd à manipuler : à part le catalogue qui a été dactylographié, la plupart des documents étaient manuscrits, et les transcriptions ne concernaient qu'une petite partie d'un corpus dont les enregistrements représentent plus de 300 bandes magnétiques .

Ensuite, ESLO n'a été que très partiellement transcrit. La transcription d'un corpus d'une telle taille constitue un défi colossal. En plus de l'absence de cadre théorique sur la transcription du français parlé – le travail théorique sur cette question débutera véritablement avec les travaux de Claire Blanche-Benveniste dans les années 1980 –, le temps de transcription peut être estimé à 20 000 heures de travail pour un coût qui dépassait de loin les moyens dont pouvaient disposer une équipe. Il n'y aura donc pas de transcription de l'intégralité du corpus.

Enfin, et surtout, l'absence d'exploitation de ce corpus peut s'expliquer par une raison épistémologique profonde : le français parlé est loin d'être un domaine légitime dans le champ de la linguistique. Cette situation constatée à plusieurs reprises notamment par Claire Blanche-Benveniste (Blanche-Benveniste et Jeanjean (1987))[8] est due à trois raisons principales. Premièrement, la linguistique en tant que discipline universitaire a *incorporé* au sein même de son organisation, la domination symbolique de la forme écrite de la langue en faisant la part belle à la grammaire normative (Bergounioux 1992). Deuxièmement, la linguistique s'est construite sur une lecture des dichotomies proposées par Ferdinand de Saussure qui a située, hors de la discipline, la description des variations en général et des formes orales en particulier. L'étude du français parlé s'est alors trouvée exclue du champ d'une science vouée à la recherche d'invariants structurés en système. Enfin, l'histoire même du français en France confirme la place attribuée à l'écrit, longtemps considéré comme seul véritable capital symbolique légitime.

2. Reconstruire ESLO1, penser ESLO2

L'apparition des nouvelles technologies et du traitement informatique des corpus va relancer des possibilités d'exploitation du corpus ESLO. Pour les raisons épistémologiques évoquées plus haut, il n'est pas étonnant que celles-ci se fassent en dehors de la France.

[7] On relève seulement deux utilisations du corpus dans le domaine du FLE : (i) la méthode anglaise diffusée sous le titre « Les Orléanais ont la parole » (livre du maître, livre de l'élève et support de cours sur bandes), et (ii) la méthode du BELC (12 cassttes).

[8] « Mais qui s'intéresse au français parlé ? […] peu de gens y voient un objet légitime d'étude (même chez les linguistes) pour bon nombre de ceux-ci la langue parlée c'est bon pour l'exotisme ; la description de la langue parlée vaut pour les dialectes et les patois du français ; elle vaut aussi pour les langues sans écritures dites "exotiques" ; mais pas pour une langue de culture comme le français. » Blanche-Benveniste et Jeanjean (1987).

On doit en effet aux universités de Louvain et d'Amsterdam, dans le cadre des deux projets ELILAP puis ELICOP, la diffusion d'un corpus informatisé[9] comprenant une partie du corpus d'Orléans transcrit, avec un étiquetage morphosyntaxique et un concordancier. La partie disponible représente près de 80 heures (900 000 mots) de transcription orthographique et une dizaine d'heures de transcription phonétique[10].

Depuis, ce corpus a donné lieu à de nombreuses exploitations et continue d'être utilisé et développé à l'université de Louvain[11].

Le CORAL, qui détenait l'intégralité des enregistrements originaux, a entrepris récemment la numérisation des bandes magnétiques à des fins de conservation et de diffusion.

Le traitement en 2005 d'un corpus vieux de 35 ans n'est pas une chose aisée et implique une véritable reconstruction, reconstruction d'autant plus nécessaire qu'elle devait répondre à un objectif de comparabilité avec une nouvelle enquête sociolinguistique à Orléans, ESLO2.

Cette nouvelle enquête, en cours d'élaboration au sein du CORAL[12], consiste à constituer un corpus suffisamment analogue à ESLO1, y compris sur le plan quantitatif, et en même temps adapté à la situation contemporaine. L'enquête portera sur 200 témoins, les enregistrements débuteront fin 2006.

Reconstruire ESLO1, c'est donc aussi penser ESLO2 avec l'évaluation de 35 ans d'évolution technologique, méthodologique et théorique. Une évaluation qui permet d'anticiper les questions de conservation, de transcription, d'annotation, de structuration des métadonnées, de multi-exploitation et surtout l'impact de ces différents choix sur l'analyse linguistique.

2.1. Quand la numérisation transforme l'objet scientifique

La conservation d'archive consiste à trouver les garanties d'une non altération du support d'origine et, à défaut – systématique à ce jour, à dupliquer l'original sur un nouveau support avec le moins de perte possible. En ce qui concerne les bandes magnétiques d'ESLO, la copie devenait indispensable – 35 ans est une durée critique pour ce type de support. Or, la numérisation ne consiste pas simplement en un changement de support. La digitalisation transforme l'objet en en facilitant la manipulation et la diffusion, mais surtout en offrant des possibilités de traitement informatique des données primaires et des métadonnées. Cependant ces traitements demandent de repenser la structure du corpus et des objets qui le composent.

[9] Dans le cadre d'un projet de recherche mené de 1980 à 1983 (Le français parlé. Banque de données automatisée ; analyse linguistique fondamentale et applications, sous la direction de Josse De Kock, Mark Debrock, Nicole Delbecque et Ellen Bas), le Département de Linguistique de la K.U.Leuven a reçu la gestion de l'ensemble de ces enregistrements, soit près de 500 heures. Ce premier projet est connu sous le sigle ELILAP (*Etude LInguistique de la LAngue Parlée*). Les responsables du projet ont voulu rendre accessibles les données sous forme informatisée. Plusieurs parties des trois corpus ont été transcrites et sont actuellement disponibles sous forme de transcriptions graphiques (± 100 heures) et phonétiques (± 12 heures) automatisées. L'ensemble des corpus compte actuellement plus d'un million de mots et peut être considéré comme constituant un échantillon représentatif de la langue parlée

[10] http://bach.arts.kuleuven.be/elicop/

[11] autour du Professeur Piet Mertens.

[12] en partenariat avec le CELITH et MODYCO

Après la première phase consacrée à la numérisation des enregistrements selon les normes en cours[13], il a fallu déterminer les opérations de traitements applicables à ces données informatiques. De fait, la chaîne de traitement élaborée doit permettre de transcrire les données primaires pour pouvoir disposer de tous les outils développés dans le domaine du TAL et de la linguistique de corpus. En ce sens la transcription doit suivre des conventions suffisamment proches des corpus écrits afin de permettre notamment un étiquetage morphologique, prosodique, syntaxique ou tout autre annotation. Nous reviendrons par la suite sur les effets de cette normalisation de l'oral pour les outils de la linguistique de corpus pour nous intéresser maintenant à un effet de la numérisation qui est souvent ignoré : le rôle des métadonnées[14].

2.2. Numérisation des métadonnées : la réapparition du locuteur

2.2.1. Traitement informatique du témoin

Numériser un corpus implique également de traiter différemment la documentation et les descriptions qui peuvent éventuellement accompagner celui-ci et qui sont, nous l'avons souligné, volontairement très riches dans le cas d'ESLO. Dans le cas d'un corpus numérique, il est aisé d'établir des relations entre les données primaires et les principes d'élaboration du corpus, la normalisation et les formalismes choisis, les techniques utilisées et de nombreuses autres informations (ou méta-informations). Or, cette documentation est notamment le lieu pour fournir de précieux renseignements sur la situation de collecte et le profil des témoins. Cette opération a été repérée comme étant fondamentale depuis le développement de la linguistique de corpus : " *La documentation doit couvrir deux volets distincts : les sources utilisés et la responsabilité éditoriale de constitution du corpus d'une part, les conventions d'annotation d'autre part*" (Habert et al.1997[15]). Récemment, le langage XML apporte une solution convaincante en séparant les données et les informations sur la structure des données, alors décrites dans l'en-tête du document (recommandations de la TEI[16]).

La gestion de ces informations souvent répertoriées sous le terme de métadonnées rend nécessaire une uniformisation du traitement comme le proposent différentes initiatives centrées sur la gestion, la diffusion et la réutilisation des corpus (EAGLES[17], OLAC[18]).

Dans le cas du corpus d'ESLO, nous disposons d'un exemple particulièrement intéressant car les métadonnées avaient déjà été répertoriées pour la publication du catalogue en 1974[19]. Or, la transformation du catalogue en une base de données offre des perspectives infinies de requêtes dans d'excellentes conditions. Ainsi, les données sociologiques ont été intégrées à des bases de données relationnelles et deviennent facilement disponibles comme champs que

[13] Les choix qui ont prévalu à la numérisation des enregistrements d'ESLO correspondent aux recommandations de l'IASA (association internationale d'archives sonores et audiovisuelles) et de l'AFAS[13] qui diffusent en France ces recommandations. Une copie droite de chaque bande a été réalisée en WAV, 44100 htz, 16 bits. Une attention particulière a été apportée aux choix de codage ouvert, de format non propriétaire, de normes pérennes et de l'évolution de la technologie (fréquence largement supérieure à la fréquence d'enregistrement initiale).

[14] Pour une présentation des méthodes de constitution de corpus oral numérique : Delais-Roussarie 2000.
[15] Habert et al. 1997, p.156.
[16] Text Encoding Initiative
[17] EAGLES, 1996, Preliminary Recommendations on Spoken texts
[18] Open Language, Archive Community
[19] Catalogue ESLO, 1974

l'on peut croiser avec des requêtes sur la transcription et l'annotation des données linguistiques.

Outre les possibilités techniques de requêtes croisées, nous avons tenu à réintroduire le locuteur dans les données primaires, continuant ici notre démarche qui a consisté à rendre indissociable la transcription à la voix du locuteur.
Le corpus reprend ainsi son statut de données situées par la réapparition du locuteur dans les corpus oraux, et par la reconstruction du profil sociologique de ce locuteur comme témoin aux caractéristiques bornées par l'échantillonnage et l'enquête sociologique.

Pour ESLO2, le travail théorique sur la sociologie applicable à une enquête linguistique débute. Cependant, la gestion des métadonnées d'ESLO1 permet d'ors et déjà de penser que non seulement une intégration en nombre de ce type de données est réalisable, mais qu'il y a ici la possibilité de rendre à la linguistique la méthodologie d'une véritable science des données attestées et situées (depuis ESLO1 l'apport de la sociolinguistique et de l'ethnométhodologie modifie la donne).

2.2.2. Reconnaissance juridique du locuteur (protection des données personnelles, propriété intellectuelle)

La reconnaissance du locuteur passe également par celle de son statut. La reconstruction du corpus d'Orléans 35 ans après a donné aussi lieu à un travail sur les aspects juridiques de l'exploitation de corpus. Ce travail a été mené parallèlement aux initiatives actuelles et notamment le groupe de travail sur le *Guide des bonnes pratiques* pour la constitution et l'exploitation des corpus oraux[20].

Le cas du corpus d'Orléans est un cas d'école : aucune autorisation n'avait été demandée à l'époque, ni sur les enregistrements, ni sur leur exploitation. Le corpus contient des données privées (nom, profession), des données sensibles (récit de vie nominatif, préférence politique, religieuse, enregistrements confidentiels au CMPP). De plus, les différentes phases d'exploitation multiplient les questions de propriétés (Essex, Orléans, Louvain,…).

A ces questions juridiques, le CORAL a apporté un certains nombres de réponses :

- les enregistrements seront anonymisés (bipage en temps réel des données personnelles, lors de la consultation)
- la structure de la base de données permet différents niveaux d'accès (toutes données pour un certain type de recherches selon une charte de confidentialité, et pour la conservation, données partielles pour une diffusion du corpus à la communauté scientifique, et données publiques pour une diffusion large).
- Le CORAL, qui a obtenu des financements de l'Etat pour ce programme, s'est engagé à suivre les règles de mise à disposition des corpus pour la communauté scientifique et pour les institutions patrimoniales. Il sera ainsi librement disponible et déposé à la BnF.

Là encore l'expérience d'ESLO1 nous as permis d'entreprendre un travail en profondeur sur l'élaboration des autorisations que rempliront les enquêtés afin de permettre le recueil d'un consentement le plus éclairé possible. Les opérations d'anonymisation et la gestion des droits de propriétés intellectuelles seront traités uniformément sur les deux corpus.

[20] Baude O (coord.) 2006,

2.3. Transcription synchronisée : la réapparition du locuteur - suite

2.3.1. La transcription synchronisée et le retour à la source

Nous l'avions déjà précisé, la difficulté la plus importante rencontrée par les initiateurs d'ESLO 1 a été l'ampleur de la tâche de transcription. Sur ce point aussi, et même principalement, l'avancée technologique bouleverse l'objet scientifique.

Depuis quelques années, alors que la manipulation du son numérique devenait très aisée (capacité de stockage, rapidité d'accès, débit suffisant pour une transmission en réseau…), des logiciels permettent la synchronisation du son et de la transcription (*Praat*, *Transcriber*, *Winpitch*, *soundedit*, etc.).
Ces innovations ont des répercussions méthodologiques importantes sur le travail du linguiste.

D'abord, les outils du traitement automatique des corpus écrits deviennent utilisables sur des données orales, qui d'un coup rattrapent 25 ans de recherche. Ensuite, avec des transcriptions alignées sur le signal sonore, l'oral devient physiquement l'objet d'étude et est systématiquement disponible en même temps que la transcription. Le retour aux données peut alors être systématique, ce qui est de nature à faciliter les procédures de vérification, étape essentielle du travail scientifique, malheureusement souvent rendue impraticable de par l'inaccessibilité des corpus.

Parallèlement, la synchronisation, qui permet l'annotation de segments temporels, offre une base de référence pour de la multi annotation et donc de la multi transcription. On peut concevoir, pour un même segment, une multitude de transcriptions, opérées dans des cadres théoriques distincts et/ou avec des granularités différentes, dont chacune répond à un besoin scientifique spécifique. Ici, la transcription n'est plus la vérité d'un chercheur (au mieux) ou d'un transcripteur, elle devient cumulative.

2.3.2. La transcription de degré 0 comme alternative à la transcription figée

Face à l'ampleur de la tâche, les choix actuels du CORAL ont été fondés sur la volonté de mettre à disposition une transcription de l'intégralité du corpus le plus rapidement possible sans que celle-ci n'implique une théorie linguistique très déterminée (même si toute transcription est une formalisation impliquant une théorie).
Nous avons conçu cette première transcription à un degré le plus proche du zéro, en lui donnant uniquement le statut d'outil de navigation au sein du corpus sonore. L'outil sélectionné a été *Transcriber* pour sa simplicité d'utilisation, sa robustesse face à des fichiers long, et sa sortie en un format de fichier XML qui nous a semblé être une garantie d'interopérabilité.

Les conventions de transcriptions ont donc été réduites au minimum. Cependant, même à ce niveau "zéro", de nombreuses questions restent présentes comme la structuration des segments et leur granularité – qu'est-ce qu'un mot ? une phrase ? un tour de parole ? –, le choix des évènements à transcrire, la gestion des chevauchements et des pauses.

Ce choix de transcription est actuellement testé sur des extraits de corpus[21]. Seule une évaluation rigoureuse des contraintes de ce choix sur les autres niveaux d'annotation (morphologique, syntaxique, prosodique, etc.) et sur les analyses linguistiques qui en découleront permet de le valider ou non.

3. Conclusion

Reconstruire ESLO1 consiste, avant toute possibilité d'analyse linguistique, à rendre le corpus disponible pour la communauté scientifique et implique l'anticipation de finalités qui n'ont pas été prévues lors de l'élaboration du projet ni même lors de l'étape de numérisation réalisée actuellement. Penser ESLO2 répond au même objectif même si la compatibilité entre deux corpus ne nous est pas apparue suffisante pour élaborer une standardisation qui n'existe pas actuellement.

Derrière ces choix méthodologiques et techniques, l'enjeu du domaine même de la linguistique pointe. Il s'agit surtout pour nous de concevoir la linguistique comme nécessairement une linguistique de données tout en ne perdant pas l'occasion de rendre à la linguistique la prise en compte de la nature sociale de la langue.

Constituer des grands corpus n'est pas une fin en soi, d'autant que ceux-ci peuvent contenir des données normées ou normalisées en masse. L'engouement actuel pour les corpus oraux, s'il constitue une chance pour la linguistique, comporte un risque majeur, celui de normaliser les données orales plutôt que saisir consciemment de la variation. De même, le recours à d'autres méthodes que la prospection des données peut paraître nécessaire. On ne doit pas exclure a priori d'autres méthodologies complémentaires : on peut aussi s'autoriser à manipuler des données et les transformer en fonction d'hypothèses falsifiables.

Le corpus du Français Fondamental représentait déjà une prise de position sur l'objet de la linguistique. L'évolution des méthodologie du traitement des données a fait apparaître de nombreuses questions qui méritent d'être abordées sans concession au moment où le champ de la linguistique s'ouvre enfin aux corpus. L'objectif de l'analyse des données d'ESLO 1&2 est de participer à ce débat.

[21] Des membres du CORAL entreprennent actuellement, à partir de différents points de vue disciplinaires (sociolinguistique, syntaxe, TAL, phonologie, pragmatique, etc.), une série de recherches linguistiques qui visent un même objet, i.e. le corpus de l'omelette. Il s'agit d'une petite sous-partie du corpus ESLO1, composé des réponses de 90 témoins à la question « Comment faites-vous une omelette ? » soit au total 120 minutes environ.
Ce mini-corpus qui offre l'avantage de l'unité thématique joue pour nous le rôle de test : il s'agit de tester sur une petite échelle l'ensemble du travail que nous nous proposons d'entreprendre sur la totalité du corpus ESLO1 mais aussi sur ESLO2, qu'il s'agisse de la transcription, ou de la faisabilité de certains types de recherches linguistiques, en passant par la structuration de la base de données.

Bibliographie

Abouda L., 2004, « Deux types d'imparfait atténuatif », *Langue française*, 142, p. 58-74,

Association française des détenteurs de documents audiovisuels et sonores (AFAS) : [http://afas.mmsh.univ-aix.fr/] voir notamment Bradley, K. (dir) *Guidelines on the production and preservation of digital objects*. International Association of Sound and Audiovisual Archives.

Baude O., Jacobson M., Tchobanov A., Walter R., à paraître, « Interopérabilité des corpus sonores : le cas des corpus en français », *Colloque international Phonological variation : the case of French*, 25-27 août 2005, Tromso.

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

Baude O. (ed), 2006, *Corpus oraux. Guide des bonnes pratiques 2006*, Paris, Cnrs éditions – Orléans, PUO.

Bergounioux G., 1992, « Les enquêtes de terrain en France », *Langue française*, 93, p. 3-21.

Bergounioux G., Baraduc J., Dumont C., 1992, « L'Etude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus », *Langue française*, 93, p. 74-93.

Blanche-Benveniste C., Jeanjean C., 1987, *Le français parlé, transcription et édition*, Paris, Didier érudition.

Biggs P, Dalawood M., 1976, *Les orléanais ont la parole : Teaching Guide and Tapescript*, (livre du maître), Londres, Longman.

Biggs P, Dalawood M., 1976, *Les orléanais ont la parole : Teaching Guide and Tapescript*, (livre de l'élève), Londres, Longman.

Blanc M., Biggs P., 1971, « L'enquête sociolinguistique sur le français parlé à Orléans », *Le français dans le monde*, 85, p. 16-25.

Calas M.-F., Fontaine J-M, 1996, *La Conservation des documents sonores*, Paris, CNRS Editions.

, *Orléans Archive*, Language center, University of Esssex, Coldchester.

A. Condamine (éd.), 2005, *Sémantique et corpus*, Paris, Hermès.

Coste D. (dir), 1984, *Aspects d'une politique de diffusion du français langue étrangère depuis 1945*, Paris, Hatier.

Delais-Roussarie E. et Durand J. (ed), 2003, *Corpus et variation en phonologie du français, méthodes et analyses*, Toulouse, PUM.

EAGLES, 1996, Preliminary Recommendations on Spoken Texts, EAG-TCWG-SPT/P, Pise, Consiglio Nazionale delle Ricerche, Instituto di Linguistica Computazionale.

Encrevé P., 1976, « Labov, linguistique, sociolinguistique », in *Labov 1976*, Paris, éditions de Minuit.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A., 1964, *L'élaboration du français fondamental (1e degré)*, Paris, Didier.

Habert, B., et al., 1997, *Les linguistiques de corpus*, Paris, Armand Colin.

Mertens P., 2002 « Les corpus de français parlé ELICOP : consultation et exploitation », in Binon, J., Piet; Elen, J., Mertens, P., Sercu, Lies (eds) (2002) *Tableaux Vivants*, Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock. Leuven, Universitare Pers.

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

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 décembre1998, *Cahiers de l'ILSL*, 10, p. 39-68.

Sinclair J., 1996, « Preliminary recommendations on corpus Typology », Technical Report, Eagles.

« Speech Annotation And Corpus Tools », A special issue of Speech Communication Volume 33, numbers 1-2, 2001, Edited by Steven Bird and Jonathan Harrington.

Wynne M., 2005, *Developing Linguistic Corpora : a Guide to Good Practice*, AHDS, http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm. Visité le 10 mai 2006.