(Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?

Olivier Baude, Céline Dugua

(Re)faire le corpus d'Orléans quarante ans après : *quoi de neuf, linguiste ?*

Olivier BAUDE
Céline DUGUA
LLL[1]

Résumé : La comparaison de deux corpus d'enquêtes sociolinguistiques réalisés à quarante ans d'intervalle permet de mettre en perspective certains aspects centraux de la constitution des données et d'interroger, par delà une description des différents choix méthodologiques et théoriques opérés, la place des données dans la linguistique de corpus.

Abstract : Comparing two corpora based on sociolinguistic studies and carried out forty years apart enabled to underline certain central questions of corpus constitution, and question the description of méthodological and theorical choices, and more generally the status of the data in linguistic corpora.

Mots clés : corpus oral, variation, transcription, liaison
Key words : oral corpora, variation, transcription, liaison

Introduction[2]

Que font les linguistes quand ils constituent et exploitent des corpus oraux ? La comparaison de deux projets d'enquêtes sociolinguistiques à des fins de constitution de corpus de référence, réalisés à quarante années d'intervalle offre l'opportunité d'aborder concrètement cette problématique.

L'enquête sociolinguistique à Orléans réalisée en 1968-71 (désormais ESLO1) avait pour objectif de fournir un corpus

[1] Laboratoire Ligérien de Linguistique, Université d'Orléans, EA3850.

[2] Les travaux présentés ont été soutenus par l'ANR *Corpus en SHS* et par la DGLFLF.

représentatif du français tel qu'il est parlé. Ce corpus relevait d'un certain nombre de choix théoriques, méthodologiques et technologiques qui étaient pour certains explicites, pour d'autres révélés lors des différentes opérations de mise à disposition du corpus et pour les derniers totalement implicites. Quarante ans plus tard, le laboratoire ligérien de linguistique de l'université d'Orléans a entrepris un double projet : diffuser largement le corpus ESLO1 dans un format correspondant aux outils de traitement des données actualisés et réaliser un nouveau corpus représentatif du français parlé à Orléans dans les années 2010 (désormais ESLO2), en prenant en compte l'expérience d'ESLO1 et l'évolution des cadres théoriques et méthodologiques de la constitution et de l'exploitation de grands corpus oraux à visée variationniste. C'est bien en effet la variation qui est au cœur de la problématique de la linguistique de corpus fondée sur l'enquête. Ainsi, dans sa préface de Labov (1976), Pierre Encrevé soulignait que « *le premier geste de la reconstruction labovienne, c'est de reposer les questions des* données *de la sociolinguistique* »[3].

Or il n'est pas évident que la constitution de grands corpus réponde à l'objectif « d'adéquation observationnelle » souhaitée par la sociolinguistique et ce pour deux raisons que nous aborderons dans cet article. Premièrement, la méthodologie de corpus implique un degré de « figement » des données d'autant plus contraignant que la masse d'informations est importante et que des outils de traitement sont requis. Deuxièmement, la reconnaissance de l'hétérogénéité des données se heurte à la nécessité de catégoriser les constituants (locuteurs, situations) de cette pratique sociale, pour borner la représentativité du corpus. Dans cet article, nous nous intéressons aux opérations de transcription et de classification sociale des locuteurs. Nous défendons l'idée que le *figement* des données inhérent au travail de constitution et d'exploitation de

[3] « Les « données » de la langue dans son usage quotidien, telle que veut l'étudier Labov, ne seront « produites » qu'au terme d'un long chemin d'aveuglette où se construit pas à pas une science de l'enquête linguistique qui est la première conquête de la sociolinguistique. » (Labov, 1976: 13).

corpus est paradoxalement une opération génératrice de variations.

1. D'ESLO1 à ESLO2

ESLO1 a été conduite à partir de 1968 par des universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère. Les buts du projet sont clairement exprimés dans le texte de présentation du catalogue des enregistrements :

> (…) dès le début il s'agissait d'autre chose que d'une simple chasse aux images sonores ; bien sûr, il fallait fixer des propos vivants, mais d'une façon systématique, afin de permettre des études fondamentales dans le domaine de la linguistique descriptive, sans lesquelles le renouveau de la pédagogie ne serait, au mieux que superficiel (Lonergan, *et al.*, 1974 : 1).

Il n'est pas anodin que le catalogue des enregistrements soit le document de référence de présentation de cette enquête. Il s'agit ici du cœur du projet : rendre disponible l'ensemble du corpus « à tout chercheur intéressé ».

Le second objectif était de constituer un corpus « sociolinguistique » :

> Selon nous une recherche sociolinguistique impliquait une étude de la langue dans sa diversité plutôt que comme un tout homogène et figé. En effet, même si on étudie un état de langue à un moment précis de l'histoire, il n'empêche qu'il offre une variété à plusieurs niveaux : différences entre les générations ; différences dialectales entre communautés ; différences entre les milieux sociaux ; différences liées aux conditions de production du discours. (Blanc & Biggs, 1971: 16)

Ces deux objectifs ont permis de construire le corpus autour du concept de « portrait sonore d'une ville » afin de croiser représentativité et variations au sein d'une communauté d'auditeurs dans un espace géographique et socioéconomique clairement défini (Bergounioux, *et al.*, 1992 : 79).

De quoi est composé ce corpus ? Selon le catalogue (Lonergan, *et al.*, 1974), il y a 487 enregistrements divisés en huit catégories générales, allant de l'entretien à la conférence en passant par des enregistrements spontanés sur les marchés ou dans la rue. L'ensemble des situations représente 315 heures d'enregistrement évaluées à 4 500 000 mots[4].

Dans sa version d'origine, le corpus comprend, outre des bandes magnétiques, un catalogue reprographié des enregistrements de 265 pages, des fiches d'identification des locuteurs, des fiches relevant les réponses au questionnaire sur les pratiques culturelles et 3365 feuillets présentant les extraits de transcriptions manuscrites ou tapuscrites. Dans les années 1980-90, une partie du corpus a été transcrit et étiqueté puis mis à disposition sur la toile dans le cadre du projet ELILAP / LANCOM[5].

En 2003, l'équipe de l'université d'Orléans entreprend la numérisation du corpus ESLO1 afin de le rendre disponible dans son intégralité. Loin d'un simple transfert de support, il s'agit véritablement de reconstruire le corpus, c'est d'ailleurs sur ce constat que le projet de réalisation d'ESLO2 s'est concrétisé autour de la volonté de maitriser l'ensemble de la chaine, de la constitution à l'exploitation d'un corpus. A terme ESLO2 comprendra plus de 350 heures d'enregistrement afin de former avec ESLO1 un corpus de plus de 700 heures et atteignant les dix millions de mots.

Au-delà d'une visée cumulative qui consisterait simplement à accroître la quantité de données pour fournir des éléments d'analyse et assurer des comparaisons avec d'autres corpus, l'enjeu des enquêtes conduites dans ESLO2 est aussi réflexif (accompagner la campagne de collecte, traiter et exploiter les données pour contribuer à la définition des normes). La mise en œuvre de cette conception implique :

- une prospective sur l'exhaustivité des usages,
- un inventaire des techniques de collecte (formats d'enregistrement et numérisation),

[4] Environ 70 % du corpus présente une qualité acoustique suffisante pour une transcription.

[5] ELILAP 1980-83 puis LANCOM 1993-2001, voir Mertens (2002).

- une politique de formation des enquêteurs et d'information des témoins afin d'intégrer dans les critères de variation celle liée à l'enquêteur,
- un recueil des données en conjonction avec le recueil des métadonnées,
- un codage et un catalogage anticipant les principales requêtes,
- une transcription avec alignement sur le signal,
- un étiquetage, avec catégorisation et lemmatisation,
- une analyse syntaxique (*parsing*), en particulier pour la co-référence anaphorique,
- une procédure d'anonymisation,
- un stockage, avec archivage et indexation,
- une procédure de mise à disposition sur la toile,
- des données partagées par interopérabilité.

Derrière ces objectifs se dresse la volonté d'élaborer ESLO2 en écho à ESLO1. Nous allons maintenant décrire deux aspects caractéristiques de cette évolution des corpus.

2. Procédures de transcription

La plus grande difficulté à laquelle ont été confrontés les auteurs d'ESLO1 a été indéniablement la phase de transcription. Face à la taille du corpus, l'équipe était démunie à la fois sur le plan technologique (le corpus n'était pas informatisé) et sur le plan théorique (les grands travaux sur la transcription n'étaient pas encore publiés). Cependant, force est de constater que, sur ce point comme sur de nombreux autres, les choix et intuitions de l'équipe ont été novateurs.

Le corpus ESLO1 a été transcrit en plusieurs étapes et à différentes époques avec des objectifs eux aussi différents. Les premières transcriptions datent du moment du recueil et des quelques années qui ont suivi. Trente-six bandes, puis cinquante-six, tout ou en partie ont été transcrites à cette époque. Les transcripteurs travaillaient alors sur papier et sur machine à écrire. Dans un deuxième temps (années 1993-2001), une partie du corpus a été repris par des chercheurs de l'Université de Louvain (Debrock, *et al.*, 2000) dans le cadre du

projet Elicop[6]. Enfin, depuis 2003, le LLL (Orléans) s'est donné pour objectif de transcrire et rendre disponible l'intégralité du corpus en y associant le son, des annotations – dont la transcription – et des métadonnées. Certes l'évolution des transcriptions sur 40 ans dépend fortement des technologies, mais c'est aussi la définition même de l'écriture de l'oral qui va être bouleversée. Dès l'origine du projet, les chercheurs ont conscience de la difficulté de la tâche de transcription qui consiste à « essayer de rendre par écrit un discours sonore [ce qui] suppose l'adoption d'un code » (Blanc & Biggs, 1971 : 19). Dans les premières transcriptions d'ESLO1, le code adopté est dit semi-orthographique :

> les écarts par rapport à l'orthographe traditionnelle sont : (1) l'absence de ponctuation, qui implique la non-délimitation des phrases, la suppression des lettres majuscules : à sa place, des signes marquant les pauses, évaluées sur une échelle de durée à trois degrés : minime, moyen, long ; (2) la notation des liaisons non évidentes, des élisions, des allongements de syllabes, des *e* normalement élidés mais prononcés dans ce cas (Blanc & Biggs, 1971 : 20).

Ces choix relèvent d'un souci fort louable : formaliser le signal acoustique (tour de parole, énoncés, mots) afin d'offrir le meilleur matériau de référence pour l'analyse. A ce stade, le son peut disparaitre et les transcriptions devenir les « données primaires » ; c'est d'ailleurs ce qui arrivera par la suite dans le projet Elicop où, sur un total de 118h30, 80 heures proviennent du corpus ESLO1 (64 entretiens, environ 900.000 mots) (Mertens, 2002). Dans ce projet, la mutualisation des corpus et la volonté de pouvoir les interroger sous forme de concordancier et de corpus étiqueté ont déterminé les types de codage utilisés, à savoir un codage orthographique et/ou phonétique, le balisage de certaines particularités de l'oral comme les pauses, les liaisons, les élisions et l'annotation de certains éléments au format SGML (Mertens, 2002).

Dans le projet du LLL, la transcription n'est plus conçue uniquement comme le préalable à une étude sur corpus

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

oraux, elle est une façon de mettre en perspective les conditions de productions des données. En ce sens, elle constitue une étape qui reflète le champ de la linguistique, les théories, l'inscription du chercheur dans son domaine, et également les « attitudes » et les représentations des transcripteurs[7]. A partir des archives sonores et des transcriptions issues du projet Elicop[8], l'équipe du LLL a transcrit l'ensemble des entretiens (151 enregistrements), les entretiens de personnalités (34 enregistrements), ainsi que d'autres situations (84 enregistrements)[9] afin de disposer d'un panorama varié des différents types d'enregistrements, dans la mesure où les qualités acoustiques permettent une transcription.

La contrainte forte qui apparait et qui détermine nos choix d'annotation est la volonté de transcrire et rendre disponible une grande quantité de paroles (environ 700h). Nos conventions de transcription ont été élaborées à partir de la comparaison des conventions des différents grands projets de corpus oraux francophones dégageant ainsi les principes communs et affinant nos particularités au regard des objectifs propres du projet. Nous avons adopté des principes de base généralement partagés à savoir une transcription orthographique qui conserve les spécificités de l'oral (amorces, disfluences, répétitions, etc.), sans usage de la ponctuation, et avec la segmentation des tours de paroles. Un élément fondamental réside dans la synchronisation entre la transcription et le signal sonore par l'ajout de jalons temporels à l'aide du logiciel Transcriber[10]. Les conventions propres au projet ont été réduites au minimum, en suivant Ochs (1979 : 44) « a more useful transcript is a more selective one ».

---

[7] Les questions liées aux attitudes des transcripteurs ne seront pas abordées ici par manque de place, voir Hriba (en cours) pour plus de précisions.

[8] Les premières transcriptions tapuscrites que nous avons récupérées sous format papier n'ont finalement pas été utilisées, en revanche, celles issues de Elicop ont été intégrées à notre procédure de transcription.

[9] Ces chiffres renvoient au nombre de transcriptions en version B (voir infra) disponibles en mai 2011, et correspondent à 230 heures d'enregistrements.

[10] http://trans.sourceforge.net/en/presentation.php

Transcrire est une tâche fastidieuse et complexe qui implique de mettre en œuvre simultanément un nombre important de compétences, notamment : écouter, segmenter en tours de parole, en questions, attribuer la parole au bon locuteur, utiliser le clavier, respecter l'orthographe, respecter les conventions, anonymiser, etc. Par ailleurs, tous les transcripteurs n'ont pas la même connaissance des normes orthographiques et de nos propres conventions, ils ont aussi un rapport à l'écrit et à la norme différent. Conscients de ces contraintes inhérentes à toute tâche de transcription de l'oral, nous mettons en œuvre une procédure dont l'objectif est de rendre explicite notre démarche afin de fournir au chercheur les outils qui lui permettront de « reconstruire » le travail de transcription.

Une manière de garder la trace de la procèdure d'élaboration de nos transcriptions consiste à conserver les trois versions produites ainsi que les versions du *Guide du transcripteur* (document qui s'étoffe au fur et à mesure pour atteindre 36 pages dans la quatrième version), mais aussi les questions posées par les transcripteurs lors de leur formation et de leur travail et les réponses apportées. Une fois la maitrise du logiciel satisfaisante, l'essentiel des questions porte sur la façon de graphier certaines particularités orthographiques ou typiques de l'oral comme les néologismes, les régionalismes, les mots issus du verlan, l'usage ou non des majuscules, etc. Les réponses à ces questions sont recensées dans un lexique qui répertorie l'ensemble des décisions prises.

Avec cette méthode des trois versions, nous évaluons le temps de transcription à 10 fois pour une première version brute, 5 fois pour une deuxième et autant pour une troisième.

La définition de ce que sont les trois versions de transcription s'est progressivement affinée. Au départ, la première était une version dite brute qui fournissait une transcription alignée au son ; cette version était ensuite relue par un deuxième transcripteur qui corrigeait les segmentations et les variations orthographiques et de conventions ; deuxième version validée enfin par un dernier transcripteur dont la tâche consistait essentiellement à vérifier l'orthographe et les conventions dans le but de rendre la forme écrite acceptable. De

plus, à chaque niveau, le transcripteur pouvait être amené à modifier des formes en raison de différences d'écoute ou de perception. Plutôt que des rajouts ou des suppressions, il s'agit essentiellement de modifications, telles que celle observée dans l'entretien ESLO1_079 (4'15), où un même segment est transcrit sous trois formes différentes dans les versions A, B, C, respectivement : « enfin reprendre à travailler après », « enfin elle recommence à travailler après », « enfin elle reprend le travail après ». Plus généralement, Hriba (en cours), à partir de neuf entretiens ESLO1 (13h d'enregistrements, environ 105 000 mots), transcrits sous les 3 versions, a ont comptabilisé le nombre de modifications apportées à une transcription dans ses passages entre les trois versions. Il en ressort que pour un entretien, en moyenne, 790 modifications sont apportées : 290 entre les deux premières versions et 500 entre les deux suivantes.

Une deuxième procédure de transcription, en phase de test, consiste à définir plus précisément en quoi consiste chacune des versions en termes de tâches. Nous avons répertorié quatre tâches principales : l'écoute, la segmentation, la transcription et l'anonymisation, dont l'importance se répartit différemment selon les versions de transcription. Clarifier les tâches et les répartir dans les trois versions de transcription devrait permettre d'être plus efficace tant au niveau du temps de travail que de la qualité des transcriptions puisque nous limitons le nombre de compétences à mettre en œuvre pour une version donnée. Ainsi, la première version s'attachera essentiellement à la segmentation (en tours de paroles, codage des questions, attribution des codes locuteurs) et à la transcription brute des passages les plus facilement audibles ; la deuxième à affiner les incertitudes d'écoute et à vérifier les conventions et règles d'orthographe ; enfin la dernière à relire précisément l'orthographe et les conventions et à anonymiser les noms de famille ainsi que les passages dits délicats. Les transcriptions rendues disponibles au grand public seront les troisièmes versions, que nous savons être provisoires ; les utilisateurs auront d'ailleurs la possibilité de proposer des corrections tant orthographiques que d'écoute.

Ce travail réflexif sur la procédure de transcription est instructif. La reconnaissance de la légitimité de l'oral a orienté les recherches vers la volonté d'avoir une transcription fidèle reliée à un cadre théorique fort (notamment depuis Blanche-Benveniste). L'analyse des différentes transcriptions des ESLO[11] et des différences dans les versions de transcription montrent que cette opération est génératrice de variations. Face à ce problème nous proposons de concevoir la transcription comme une simple annotation de bas niveau qui permet de faciliter l'accès à la source sonore. Pour contrer les biais importés par cette annotation, il convient de lui redonner son rôle (celui de simple outil) et de la documenter avec beaucoup de précisions.

La transcription n'est pas la seule opération qui « instrumentalise » véritablement un corpus. La construction de la représentativité des données est également au cœur du travail du linguiste. Nous allons voir que là aussi il est nécessaire de porter une attention particulière aux effets des pratiques scientifiques qui sont loin d'être toujours explicites.

3. Représentativité des données

Le travail le plus novateur d'ESLO1 a sans conteste été l'approche sociologique de la représentativité des données. Cette question a été abordée selon deux axes : l'échantillonnage des locuteurs et la diversité des situations de communication enregistrées. Il est étonnant de constater que le deuxième axe a été traité dans un flou théorique relativement important voire avec une certaine naïveté alors que le premier révèle ce que la sociologie et l'enquête statistique portaient de plus rigoureux. Sur ces deux axes les choix d'ESLO2 sont sensiblement différents.

Le premier choix de l'équipe d'ESLO1 a été de limiter « pour des raisons d'ordre théorique et pratique, pour écarter aussi des variables incontrôlables » (Blanc & Biggs, 1971 : 16-17) le nombre de variables recherchées dans le cadre d'un

[11] Les choix de transcription dans ESLO2 sont également ceux du LLL dans sa nouvelle transcription d'ESLO1.

corpus sociolinguistique. Le choix s'est alors porté sur « le portrait sonore d'une ville ».

> C'est une communauté d'auditeurs qui est construite, autant qu'une communauté de locuteurs, à notre connaissance pour la première fois en France (…) On ne cherche pas cet individu mythique, l'Orléanais moyen (Blanc & Biggs, 1971 : 23).

Comment ne pas voir dans ces choix une similitude d'approche avec la linguistique variationniste se développant en France à la même époque (Labov, 1976), qui situe la langue du côté de la réception et qui lui confère de par sa nature sociale des variations instituées socialement ? Si le choix d'une délimitation sociogéographique d'une ville paraissait pertinent à la fin des années soixante, il est apparu aux auteurs d'ESLO2 qu'il convenait d'avoir, 40 ans plus tard, une définition des contours de la ville qui accepte des locuteurs vivant ou travaillant dans l'ensemble des communes limitrophes.

La collectivité choisie, se pose alors la question de l'élaboration de l'échantillon. Pour ESLO1 le cadre théorique sociologique est clairement celui des enquêtes statistiques de l'INSEE :

> Ce sont les services de l'INSEE, qui sur des instructions des membres de l'équipe de l'enquête, ont procédé au tirage au sort de six cents témoins répartis également entre six catégories socioprofessionnelles (Blanc & Biggs, 1971 : 17).

Les autres critères sont le sexe et l'âge. Cette méthodologie offrait clairement la possibilité d'analyses de la co-variation en croisant stratification sociale et variations linguistiques.

ESLO1 témoigne néanmoins des prémices d'un changement théorique en sociologie. Ainsi Alix Mullineaux (Mullineaux & Blanc, 1982) proposa une échelle (dorénavant échelle AM) qui complète les critères de l'INSEE par le diplôme et l'âge de fin d'étude des témoins. Cette nouvelle grille comprend cinq agrégats (de A à E). Or cette première étape, qui annonçait un travail conjoint avec le centre de Sociologie Européenne développant alors les travaux de Pierre

Bourdieu autour du capital culturel et de la distinction, est restée inachevée. Les entretiens furent toutefois complétés par un questionnaire sociolinguistique destiné à capter les pratiques et les représentations des locuteurs en ce qui concerne la langue, puis par un questionnaire sur l'ensemble des pratiques culturelles.

Cette approche était présentée comme particulièrement prometteuse dans la préface du catalogue de 1974.

> L'échelle de catégories socio-culturelles construite par Alix Mullineaux constitue une tentative de classement de la population française en fonction des paramètres de mobilité sociale et de niveaux de culture, et par là, marque un pas important vers l'élaboration des échelles proprement sociolinguistiques indispensables à la nouvelle discipline. (Lonergan, *et al.*, 1974 : 1)

A notre connaissance ce travail, pourtant central dans ESLO1 n'a pas été poursuivi et l'échelle AM qui visait l'amélioration de la catégorisation des témoins par le croisement des critères socioéconomiques avec les critères de niveau scolaire et les critères de capital culturel (dont linguistique) n'a pas été reprise.

En revanche, cette perspective apparait clairement dans l'élaboration d'ESLO2. En premier lieu le bilan de l'échec de la représentativité des critères de l'INSEE dans ESLO1 a été pris en compte. En effet, sur les 600 personnes sélectionnées au hasard, seules 144 ont accepté de participer à l'enquête et bien évidemment selon une répartition déséquilibrée.

L'échantillon d'ESLO2 a été conçu sur la base des critères de l'INSEE couplés à ceux de l'âge et du sexe, simplement comme point d'entrée afin d'orienter la sélection des 150 personnes qui participeront à des entretiens. La mise en place des entretiens a, quant à elle, bénéficié de nouvelles théories développées depuis les années 1970 :

- les développements de l'analyse de la conversation en linguistique et de l'ethnométhodologie qui se construisent sur une opposition frontale entre données provoquées et non provoquées par le chercheur ;

- les recherches en sociologie et en anthropologie sur les techniques d'enquête (Beaud & Weber, 1997 ; Bourdieu, *et al.*, 1968) ;

- les travaux sur la linguistique des genres (Biber, *et al.*, 1998) et sur une typologie des productions liée à une typologie des situations de communications (Koch & Oesterreicher, 2001) qui restreignent les entretiens à un contexte de production linguistique particulier ;

- les travaux sur les effets du dispositif technologique (qualité et discrétion du dispositif d'enregistrement, traitement des données numériques et même développement de la vidéo favorisant les travaux en linguistique interactionnelle (Goodwin, 1994 ; Mondada, 2006).

La méthodologie de l'enquête que nous ne développerons pas davantage ici s'appuie sur la nécessité de produire des entretiens qui permettront, après analyses, de procéder à une classification sociologique des locuteurs. Pour cela, l'équipe s'est tout d'abord appuyée sur un riche « portrait socio-économique de territoire[12] » rédigé par l'INSEE et comprenant 75 cartes illustrant des informations obtenues lors du recensement de 1999. Ensuite le mode d'approche des locuteurs et la « menée » de l'entretien ont été particulièrement étudiés afin de faciliter le recueil d'informations sur les itinéraires de vie et les pratiques culturelles des locuteurs tout en favorisant une parole dans un style le moins formel possible. Ainsi, les entretiens favorisent des discussions libres autour de la vie quotidienne (le quartier, les commerces, les loisirs, etc.).

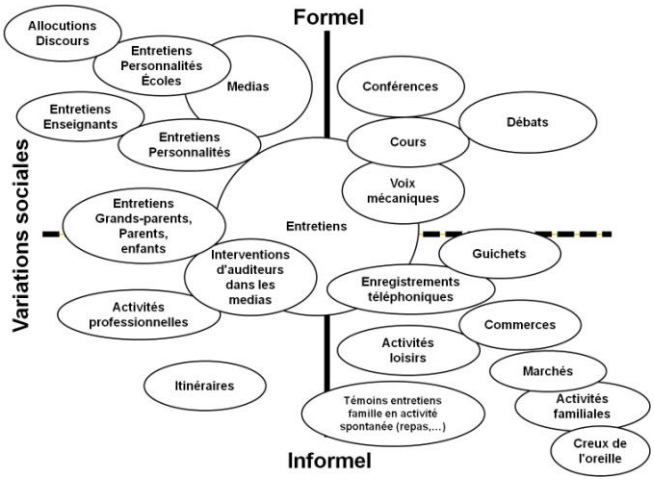
Il est alors clair que les entretiens de l'ESLO2 ne formeront pas un échantillon sociologiquement équilibré mais plutôt une archive – ou un « réservoir » – qui permettra de sélectionner à l'aide d'outils statistiques grossiers (INSEE) mais aussi à l'aide de descripteurs fins après une analyse de contenus réalisée ultérieurement, un véritable corpus à des fins d'études linguistiques.

La représentativité des données a été également abordée dans ESLO1 selon l'axe de la variation diaphasique. La variation diaphasique a cependant bénéficié d'un traitement

[12] Portrait de Territoire INSEE : 75 cartes d'analyses issues du recensement.

bien moins rigoureux que la variation diastratique mais néanmoins avec une forte intuition au regard de l'évolution des théories linguistiques (typologie des genres et linguistique interactionnelle) qui s'est développée par la suite. Dans ESLO1 cette variation a été appréhendée en enregistrant une quinzaine de témoins dans diverses situations (au téléphone, sur leur lieu de travail, en réunion, lors d'un repas de famille, etc.) et par des enregistrements en micro-cachés « au hasard des rencontres, dans la rue dans les magasins, etc. » (Blanc & Biggs, 1971 : 19). Outre les entretiens, deux autres situations avaient été clairement déterminées : l'enregistrement de tables rondes ou/et débats / conférences et des entretiens au CMPP. Pour des raisons techniques liées à la qualité acoustique mais aussi pour des raisons déontologiques et juridiques (micro-cachés) beaucoup de ces enregistrements posent des problèmes de transcription et d'exploitation. Cette partie du corpus, passionnante dans les perspectives recherchées, reste donc quasiment inexploitée.

Pour ESLO2, nous avons réparti un ensemble d'une vingtaine de modules selon deux axes : l'un prend en compte le degré de formalité de la situation et l'autre celui de la classification sociale des locuteurs :



Dans cette architecture du corpus la place des entretiens est relativisée dans une nouvelle perspective de typologie des genres de discours affinée.

La définition des « genres » repose également sur un ensemble de descripteurs contextualisant les productions linguistiques enregistrées. Afin de préserver l'accessibilité de ces descripteurs, il convient de développer des formats de codage et de catalogage. Il y a donc un enjeu à considérer ces métadonnées comme des éléments de description des données linguistiques et non simplement en termes de documentation des sources. Elles doivent permettre d'expliciter la démarche du chercheur en proposant une description fine de ses choix théoriques « encapsulés » dans des choix techniques. Si dans ESLO1 les « métadonnées » se résumaient en un catalogue constitué de fiches tapuscrites (Lonergan, *et al.*, 1974), les ESLOs bénéficient des expériences actuelles de communautés scientifiques (OLAC et DUBLIN-CORE pour le catalogage, TEI pour le codage, pour ne citer qu'elles) qui ne permettent cependant pas de structurer toutes les « données situantes » utilisées par les équipes de recherche à l'origine des deux enquêtes. Néanmoins, ces métadonnées ont été intégrées à une base de données qui permet de faire des requêtes sur le corpus en croisant les recherches sur les transcriptions et les descripteurs des locuteurs et des situations.

Ces choix de constitution et d'exploitation des corpus ESLO ont-ils un réel impact sur le potentiel d'analyse du corpus ? Afin de répondre à cette question nous nous sommes livrés à une analyse test sur le phénomène de la liaison.

4. L'exemple de la liaison

Depuis les années 1970 et tout au long des 40 ans qui se sont écoulés, la liaison a fait l'objet d'études essentielles, dans des domaines variés de la linguistique : phonologie, sociolinguistique, étude de corpus, acquisition, pour n'en citer que quelques-uns. Une des plus fameuses recherches menées sur ESLO1 porte sur la réalisation des liaisons selon une approche à la fois phonologique et sociolinguistique (De Jong, 1988, 1994). Il s'agit d'un travail descriptif d'étude de corpus qui vise à rendre compte des fréquences objectives de réalisation des liaisons, en fonction des catégories grammaticales, du lexique et des caractéristiques sociologiques

des locuteurs. En se limitant à certaines formes verbales, par exemple la forme « est », De Jong (1994) note un effet de la catégorie socio-économique, un effet du sexe (les femmes réalisent davantage la liaison) et un effet de l'âge (les témoins plus avancés en âge font davantage la liaison). Pour d'autres formes verbales, comme les auxiliaires de mode, les résultats diffèrent. Sur cet exemple en particulier, seul l'effet de l'échelle AM est significatif.

Nous avons sélectionné un sous-corpus afin de le confronter aux résultats généraux sur l'usage de la liaison analysés par De Jong. Ce sous-corpus est représentatif de la variété des situations composant ESLO1 (7 des 8 catégories sont représentées, de la prise de contact au repas de famille en passant par des visites professionnelles et des entretiens) et de la variété des locuteurs (de A à D sur l'échelle AM), en accordant une priorité aux enregistrements recoupant ces deux axes de variations – ainsi un même locuteur est enregistré dans 5 situations différentes. Les enregistrements ont une durée qui varie de 1 à 89 minutes pour un total de 11h08, soit environ 100 000 mots. A partir du travail de De Jong (1994), nous avons sélectionné 4 contextes lexicaux pouvant entrer en contexte de liaison (*est*+X, *sont*+X, *ont*+X, *quand*+X), des formes repérées comme fréquentes dans son corpus, qui proviennent de catégories grammaticales variées et qui présentent des taux de réalisation de la liaison relativement contrastés : une forme du verbe *avoir* après laquelle la liaison est peu réalisée (*ont*+X : 8.7% de liaisons réalisées), un « complémenteur » (terminologie utilisée par De Jong) après lequel la liaison est très fortement réalisée (*quand*+X : 96.3%) et deux formes du verbe *être* présentant un taux de réalisation moyen (*est*+X et *sont*+X, respectivement 69% et 46%). Dans notre sous-corpus, nous avons relevé 985 occurrences de liaisons possibles dans ces contextes. Une première analyse confirme les grandes tendances décrites par De Jong : les liaisons sont fortement réalisées dans le contexte *quand*+X (95,7%), moyennement dans le contexte *est / sont*+X (64%) et faiblement dans le contexte *ont*+X (20%).

Toutefois les variations entre les différents locuteurs sont fortes : *quand*+X de 90% à 100% ; *est / sont*+X de 36% à

100% et *ont*+X de 17% à 33%. Nous confirmons également qu'en groupant l'ensemble des situations nous constatons une co-variation entre l'échelle AM et le taux de réalisation des liaisons (de A : taux de liaison plus fort à D : taux de liaison plus faible). De même les variations pour un même locuteur selon les situations de communication sont extrêmement contrastées.

Ainsi, le locuteur 1134 (entretien 024) réalise 100% des liaisons après *est / sont* pendant l'entretien mais ce taux descend à 50% et même 19% en situation informelle lors d'enregistrements avant et après l'entretien[13]. Le locuteur BA725 (entretien 001) réalise quant à lui 75% de liaisons après *est / sont* en entretien et un taux relativement proche de 67% hors entretien. Enfin, un troisième locuteur, 1268 (entretien 029), réalise 78% de liaisons en entretien dans le même contexte et 75% hors entretien. Ces chiffres, éclairés par les informations sur le profil sociologique des locuteurs ramené à l'échelle AM, trouvent une explication compatible avec les théories développées en linguistique variationniste. La locutrice 1268, une jeune femme étudiante issue de la grande bourgeoisie et classée A par l'échelle AM, présente le taux de liaison le plus stable. Quant au locuteur 1134, vendeur de 48 ans ayant arrêté ses études à 13 ans (D sur l'échelle AM), il produit la plus forte variation avec un taux particulièrement faible dans la situation d'après entretien.

Le cas du second locuteur (BA725) est particulièrement intéressant car bien que d'un profil proche de 1134 – boucher de 57 ans ayant arrêté ses études à 14 ans (également D sur l'échelle AM) – son taux de liaison réalisée est relativement stable selon les situations et donc proche des productions de 1268 (A sur l'échelle AM). Toutefois, une analyse du contenu des entretiens permet de pondérer la classification AM. Ainsi 1134 était « charron », « rêvait d'être boulanger », a un « fils militaire et une fille mariée », « ne prend pas de vacances sauf

[13] Il est vraisemblable que les enquêteurs n'attiraient pas l'attention du locuteur sur le fait que l'enregistreur fonctionnait avant et après l'entretien. Outre le fait que cette technique était utilisée à l'époque par W. Labov, la qualité acoustique confirme cette hypothèse.

la pêche », a été « une fois au cinéma en 17 ans » et ne connait pas « le dictionnaire utilisé par [son] enfant ». BA725 souhaite devenir « gérant de plusieurs boucheries », il aime « la lecture et la musique », compte « visiter des musées quand [il sera] à la retraite »… Évidemment, ces quelques informations ne peuvent résumer le contenu de l'entretien ; cependant même si ces deux locuteurs sont classés avec le même code de CSP, et le même code de l'échelle AM (même niveau d'études et même âge de fin d'études), on repère aisément deux trajectoires différentes. Ces trajectoires sont cohérentes avec l'habitus linguistique de ces locuteurs, habitus qui entraine 1134 à adapter son taux de liaison au marché linguistique (entretien *vs* hors entretien) alors que BA725 possède la même stabilité que la locutrice 1268. Il ne s'agit ici que d'hypothèses qui ne peuvent être totalement étayées. Les faiblesses dans l'architecture d'ESLO1, le manque d'informations sur les cadres théoriques et surtout sur la méthodologie mise en œuvre par l'équipe de chercheurs tout comme la difficulté à adopter une démarche réflexive sur les effets de figement des données ne permettent pas d'explorer totalement ces pistes d'analyses.

Conclusion

Est-ce que les corpus ESLO1 et ESLO2 répondent aux objectifs de reconstruction de la linguistique à partir des données, tel que proposé dans le programme de la sociolinguistique ? D'une manière générale est-ce une bonne idée que de faire des corpus ? Ces questions méritent d'être abordées à l'aide d'analyses linguistiques. Ainsi, ce que nous apprend l'étude comparée de deux corpus à quarante années d'intervalle c'est surtout que le dialogue entre sociolinguistique et linguistique de corpus est des plus fructueux sous réserve que toutes les étapes de la collecte à l'analyse soient abordées avec la même rigueur et le même souci de réflexivité.

L'exemple d'analyse des variations du taux de liaisons réalisées selon le profil sociologique du locuteur et la situation (marché linguistique) illustre la difficulté pour le linguiste à manipuler les données d'un corpus. Si un « tamisage » (transcription et codage du signal, classification socio-

économique et échelle AM, catégorisation des situations) permet d'organiser les données afin de dégager des pistes d'études sous la forme de grandes tendances, l'analyse doit pouvoir s'appuyer sur une observation minutieuse des productions linguistiques en situation. Pour cela, il faut que l'élaboration du corpus favorise et anticipe ce retour aux données primaires non dégradées. C'est bien ce que visent les principaux choix de constitution d'ESLO2.

Références bibliographiques

Site ESLO : http://eslo.in2p3.fr

Beaud S. & Weber F. (1997). *Guide de l'enquête de terrain: produire et analyser des données ethnographiques*. Paris : La Découverte.

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

Biber D., Conrad S. & Reppen R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge : Cambridge University Press.

Blanc M. & Biggs P. (1971). « L'enquête socio-linguistique sur le français parlé à Orléans », *Le français dans le monde* 85 : 16-25.

Bourdieu P., Chambordon J.-C. & Passeron J.-C. (1968). *Le métier de sociologue*. Paris : Mouton de Gruyter/Bordas.

De Jong D. (1988). *Sociolinguistic aspects of French liaison*, Academisch proefschrift. Amsterdam : Vrije Universiteit Amsterdam.

De Jong D. (1994). « La sociophonologie de la liaison orléanaise », in C. Lyche (éd.) *French Generative Phonology: Retrospective and Perspectives*. Salford : ESRI, 95-129.

Debrock M., Mertens P., Truyen F. & Brosens V. (2000). *ELICOP, Etude LInguistique de la COmmunication Parlée : Constitution et exploitation d'un corpus de*

français parlé automatisé. K.U.Leuven : Departement Linguïstiek.

Goodwin C. (1994). « Recording human interaction in natural settings », *Pragmatics* 3 : 181-209.

Hriba L. (en cours). *Identification automatique des locus de variation dans un corpus de français parlé*, Thèse de doctorat. Université d'Orléans, Orléans.

Koch P. & Oesterreicher W. (2001). « Langage parlé et langage écrit », in G. Holtus *et al.* (eds) *Lexikon der romanistischen Linguistik*. Tübingen : Max Niemeyer Verlag, 584-627.

Labov W. (1976). *Sociolinguistique*. Paris : Editions de Minuit.

Lonergan J., Kay J. & Ross J. (1974). *Etude sociolinguistique sur Orléans, catalogue des enregistrements*. Colchester : Multigraphié.

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

Mondada L. (2006). « Video recording as the preservation of fundamental features for analysis », in H. Knoblauch *et al.* (éd.) *Video Analysis*. Bern : Lang, 51-68.

Mullineaux A. & Blanc M. (1982). « The problems of classifying the population sample in the socio-linguistic survey of Orléans (1969) in terms of socio-economic, social and educational categories », *Review of Applied Linguistics* 55 : 3-37.

Ochs E. (1979). « Transcription as theory », in E. Ochs & B. Schieffelin (eds) *Developmental pragmatics*. New York : Academic Press, 43-72.