“ Les ESLOs, du portrait sonore au paysage digital ”

Olivier Baude, Céline Dugua

Les ESLO, du portrait sonore au paysage digital

Baude Olivier, Céline Dugua
Laboratoire Ligérien de Linguistique, UMR 7270

Résumé : Cet article souhaite porter un regard réflexif sur un projet scientifique de constitution et d'exploitation d'un grand corpus de français parlé, les *Enquêtes sociolinguistiques à Orléans*, né à l'aube de la sociolinguistique et qui se développe au tournant méthodologique et épistémologique des *digital humanities*. Quels objectifs ? Quelles données ? Quels traitements ? Ce sont les questions qui guident la réflexion proposée ici afin d'apporter une contribution à l'élaboration de nouvelles pratiques scientifiques dans une perspective variationniste contemporaine.

Abstract : This article is an analysis of the constitution and the exploitation of a large corpus of spoken French: *Les Enquêtes sociolinguistiques à Orléans* (ESLO). This corpus has been created from the beginnings of sociolinguistics and now it evolves with digital humanities, methodological and epistemological specificities. Which objectives? Which data? Which analysis? These are the questions that guide our thinking in order to contribute to the elaboration of new scientific practices in a variationnist perspective.

Mots clés : Sociolinguistique, corpus, linguistique variationniste, digital humanities.
Key words : sociolinguistic, corpora, variationnist linguistic, digital humanities.

Les Enquêtes sociolinguistiques à Orléans (dorénavant ESLO) forment un grand corpus oral de plusieurs millions de mots. Ces corpus ont été réalisés à deux époques importantes de la linguistique contemporaine. La première (ESLO1), élaborée à la fin des années soixante, accompagne la naissance d'une sociolinguistique urbaine fondée sur un grand corpus d'enquêtes, et la seconde (ESLO2), commencée au début des années 2000, a profité du tournant numérique produit par les *Digital Humanities* en sciences humaines et sociales. Résolument ancrées dans le courant de la sociolinguistique et de la linguistique variationniste, les ESLO forment le socle d'études sur le français parlé à Orléans dans une perspective qui place les données au coeur d'études sur la nature sociale de la langue.

Cet article vise à décrire le travail réalisé depuis une dizaine d'années par l'équipe du projet des ESLO en le confrontant à ses cadres théoriques et méthodologiques. Après avoir abordé brièvement l'ancrage sociolinguistique du statut des données et le périmètre du français parlé, nous présenterons le travail réalisé afin de faire de ces corpus un « objet scientifique disponible » et situé.

1. Sociolinguistique et corpus

La notion de corpus croise différentes approches parfois relativement éloignées selon qu'on se situe dans une perspective de linguistique de terrain ou de linguistique informatisée. Elle prend néanmoins un sens bien plus défini dans le cadre du programme de la sociolinguistique tel qu'il a été établi dans la seconde moitié du vingtième siècle.

1.1. Nature sociale de la langue

La sociolinguistique s'est fondée sur une relecture pertinente de définition même de l'objet de la linguistique et sur la volonté de couvrir l'ensemble du domaine.

> *Pour Labov, la sociolinguistique n'est pas une des branches de la linguistique, et pas davantage une discipline interdisciplinaire : c'est d'abord*

> *la linguistique, toute la linguistique - mais la linguistique remise sur ses pieds. Elle se fonde sur l'ambition de remplir dans sa totalité le programme que la linguistique se donne dans sa définition moderne – et de l'outrepasser du seul fait de ne pas réduire son objet*. (Encrevé, 1976 : 9)

Dans cette perspective, la sociolinguistique définit la langue comme étant *partie prise* et *partie prenante* d'un social qui ne peut se réduire à un trésor collectif. Si le social est divisé et lieu de luttes et d'enjeux qui le structurent, la langue en porte, dans sa nature même, les caractéristiques qui font de la variation le principe même de celle-ci :

> *Une partie fondamentale des variations présentées par les paroles individuelles est elle aussi « instituée socialement », et par là même gouvernée par des règles : elle fait partie du système de la langue. Elle trouve normalement sa place dans la « linguistique interne » telle que la définit le CLG : « Est interne tout ce qui concerne le système et les règles (…) est interne tout ce qui change le système à un degré quelconque ».* (Encrevé 1976 :11-12)

Cette conception de la variation comme composante inhérente de la langue a une incidence directe sur la définition de l'objet d'étude sur lequel les linguistes doivent se pencher. Si les variations linguistiques sont à étudier au sein du domaine de la linguistique interne, la langue est bien le lieu où productions linguistiques et marché linguistique sont étroitement liés selon une « grammaire de la réception » qui situe la langue, comme le faisait déjà Saussure, dans le circuit de la parole :

> *Ainsi la langue d'un sujet, contrairement au sujet commun, ce n'est pas la langue qu'il parle, c'est la langue qu'il entend. Or que reçoit l'oreille d'un sujet parlant : très précisément ce que la sociolinguistique veut enregistrer et que la linguistique actuelle refuse d'écouter, les multiples paroles dont l'ensemble hétérogène*

> *arrivera à former la langue de la communauté* (Encrevé 1976 : 7).

Ainsi la communauté linguistique doit être saisie en tant qu'organisation concrète structurée et structurante des dynamiques sociales. C'est bien au cœur de celles-ci plutôt que dans une recherche illusoire d'une langue stabilisée au sein d'une communauté homogène, qu'il faut aller observer la langue afin d'obtenir l'adéquation observationnelle première que Chomsky lui-même réclamait.

> *Au total, c'est dans le caractère intrinsèquement social de la langue, dans l'intimité du lien entre langue et communauté linguistique socialement qualifiée que Weinreich, Labov et Herzog (1968) voient la source première et le moteur du changement linguistique. La communauté linguistique rappellent-ils, est une organisation sociale concrète. Elle est donc, ex definitio, profondément hétérogène, divisée, hiérarchisée, structurée par des dynamiques sociales antagoniques. La variation et l'hétérogénéité linguistique d'une part, la variation et l'hétérogénéité sociale de l'autre, ne sont alors que les deux aspects du même réel social. C'est ainsi parce qu'il n'existe jamais de communauté homogène parfaitement stable qu'il n'existe jamais de langue homogène parfaitement invariante et stable.* (Laks 2013 : 41)

Là encore, la langue ne peut se définir en dehors d'un réel social qu'il convient d'appréhender pour toute étude sur la langue. Selon Bourdieu, l'expression linguistique résultait d'une production émanant d'un habitus linguistique confronté à un marché linguistique (Bourdieu 1984 : 121). Il en résulte que l'acquisition du langage met en jeu des intériorisations socialement réglées. Ainsi comme le souligne Encrevé :

> *Aussi la grammaticalité est-elle toujours de nature sociale quant à son origine concrète pour un sujet : elle est toujours reçue et acquise assortie de sanctions sociales, dont la nature et*

> *l'importance varient avec le marché de la langue en cause – corrections, reprises, réprimandes dans la famille ; rire, moquerie de la part des égaux pour les dialectes dominés ; sanctions du marché scolaire, du marché matrimonial, du marché du travail pour les dialectes dominants.* (Encrevé 1976 : 7-8)

Il est alors aisé de concevoir le changement linguistique comme un processus résultant d'une lutte au sein de l'hétérogénéité des pratiques linguistiques évaluées socialement. La boucle est bouclée, de l'acquisition du langage au changement linguistique, la sociolinguistique offre un cadre théorique où la nature sociale de la langue est maintenant clairement définie. Cette définition de l'objet de la linguistique par la sociolinguistique se concrétise en premier lieu, et de manière centrale, autour de la question des données.

1.2 Sociolinguistique et données

En effet, définir la langue comme un fait social, nécessite de l'observer comme une pratique socialement située. C'est donc au sein même de l'activité sociale qu'elle devient appréhendable :

> *Partie structurée d'un tout qu'elle structure, la langue, en effet, n'est jamais « donnée ». Les « données » de la langue dans son usage quotidien, telle que veut l'étudier Labov, ne sont « 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* (Encrevé 1976 : 13).

Pour la sociolinguistique, il ne s'agit pas d'une simple question méthodologique qui déterminerait l'observation des données comme une étape préliminaire à l'analyse scientifique, bien au contraire la définition même des données et des conditions de leur production sont au cœur du travail du linguiste. La

première incidence concerne le périmètre des données linguistiques. Comme le souligne Laks (2013), on ne peut concevoir d'analyser des données linguistiques orphelines de l'habitus du locuteur et du marché qui structure ses productions :

> *Observer la variation dans sa systématicité et rendre compte de l'hétérogénéité comme étant structurée impose évidemment d'adopter une méthodologie adéquate. On sait en effet que décontextualisée, l'observation détruit la systématicité des phénomènes variables et les fait paraitre erratiques. Observer les faits linguistiques hors de l'écosystème social qui les conditionne détruit en effet tout ce que la pratique doit précisément à son caractère pratique. C'est la raison pour laquelle l'analyse de la variation systémique commence nécessairement par une réflexion critique sur les observables.* (Laks 2013 : 36)

Dans les années soixante-dix, la réflexion sur la place des données a entrainé une véritable science de l'enquête linguistique pour laquelle les avancées de la sociologie à la même époque, depuis Bourdieu, Chamboredon et Passeron en 1968 jusqu'à Beaud et Weber en 1997, ont été déterminantes en ce domaine. Parallèlement et parfois simultanément à l'apport de la sociologie de l'enquête, la naissance du domaine de l'analyse de conversations et les études sur les données « naturelles » ou plus justement sur les données issues de « situations non provoquées par le chercheur » sont également des éléments essentiels du développement de la science de l'enquête linguistique.

Enfin, le troisième domaine constitutif de cette démarche méthodologique et théorique provient de la linguistique de corpus dans son versant « informatique et traitement automatique du langage ».

1.3 Données et posture du chercheur

Dans cette perspective la place des données devient prédominante et le travail du linguiste ne peut s'affranchir d'une démarche réflexive sur la méthodologie de constitution et d'exploitation des données. Il lui revient alors de rendre explicite ses motivations scientifiques, sa méthodologie de collecte, la description des données et le traitement de celles-ci (Habert, 2005). C'est alors une véritable posture qui se profile sur la base d'une confrontation scientifique qui doit rendre possible la disponibilité des données, y compris pour un retour évaluatif ou contrastif, leur interopérabilité et leur description fine. En outre cette posture ne peut s'affranchir d'une réflexion éthique et juridique (Baude, 2006) sur les données, les locuteurs et le terrain non exempts d'enjeux sociaux.

Il s'agit donc de définir une conception de la sociolinguistique et par-delà de la linguistique, à partir de la relation de cette discipline aux données, nécessairement variationnistes et situées. Ceci nécessite que le linguiste sache ce qu'il fait (Gadet 2007) dans la continuité d'une évolution méthodologique et théorique d'une science de l'enquête à une science du corpus.
Les Enquêtes sociolinguistiques à Orléans, qui se concrétisent par un ensemble de deux corpus réalisés à quarante années d'intervalle, offrent l'opportunité d'évaluer, à partir de projets concrets, le cadre de ce positionnement.

2. Le français ordinaire

2.1 La recherche du français parlé

ESLO1 a pour origine un projet à finalité didactique. L'équipe constituée à la fin des années soixante autour de Michel Blanc avait comme objectif de réaliser une méthode d'enseignement audiovisuelle du français langue seconde à partir de documents authentiques. Celui-ci est clairement défini dans un court article paru en 1971 (Blanc & Biggs). A « une époque où le rôle essentiel de la langue parlée dans l'enseignement d'une langue étrangère » venait d'être acquis, il a fallu « constituer un

ensemble cohérent de matériaux vivants, rassemblés de manière systématique » valable « à la fois pour l'application pédagogique et pour la recherche sur la langue parlée ». Partant du constat qu'une collection ordonnée de documents de ce type n'était disponible, l'équipe a entrepris de collecter un vaste corpus représentatif du français parlé à partir d'une enquête ciblée sur une ville « moyenne » française exempte de caractéristiques trop marquées.
La démarche a d'emblée été résolument ancrée dans le champ de la sociolinguistique et la variation fut au cœur du travail de définition de la représentativité du corpus :

> *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).

Cette prise en compte de la diversité n'exclut pas, bien au contraire, la recherche d'une langue partagée par une communauté linguistique. C'est ainsi que le projet s'est orienté vers la réalisation du portrait sonore de la ville d'Orléans. Il s'agissait d'observer et de capter à un moment précis, dans un lieu restreint, la dynamique des pratiques linguistiques partagées par les habitants d'une cité. Le corpus est donc constitué d'une collection d'entretiens de locuteurs socialement situés et catégorisés, mais aussi d'enregistrements variés donnant accès au « français parlé dans une ville moyenne par la population de la ville à une époque précise » (Blanc & Biggs 1971).

2.2 La découverte du français entendu

La grande originalité pour l'époque et le parti pris très fort choisi par l'équipe a été de définir les pratiques linguistiques communes non pas par les productions de locuteurs types mais par l'hétérogénéité des pratiques linguistiques entendues dans la ville. Comme le soulignent Blanc & Biggs « 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). On est ici dans la même perspective de la sociolinguistique que celle défendue par Encrevé quelques années plus tard quand il reprend l'affirmation de Saussure selon laquelle la langue comme objet de la linguistique se situe dans le circuit de la parole, pour préciser immédiatement que

> *pour Saussure la langue est entièrement, et exclusivement, du côté de l'audition, de la réception : on peut la (la langue) localiser dans la partie déterminée du circuit (de parole) où une image auditive vient s'associer à un concept ; c'est par le fonctionnement des facultés réceptives et coordinatives que se forment chez les sujets parlants des empreintes qui arrivent à être sensiblement les mêmes pour tous. Ces deux points sont manifestement reliés : seule l'audition met le sujet en contact avec la masse parlante. Ainsi la langue d'un sujet, contrairement au jugement commun, ce n'est pas la langue qu'il parle, c'est la langue qu'il entend* (Encrevé 1977 : 6).

Nous le verrons dans le chapitre consacré à l'architecture des corpus des ESLO, ce cadre théorique et ses incidences méthodologiques apportent une très forte identité à l'ensemble du projet.

2.3 La linguistique du français parlé d'ESLO1 à ESLO2

Entre les deux enquêtes ESLO1 et ESLO2, la linguistique française a bénéficié des très précieux travaux de Blanche-Benveniste et de l'école du GARS sur la description du français

parlé. Ces études, principalement grammaticales, ont incontestablement marqué le champ de la discipline. Or, comme ces travaux du GARS reposent essentiellement sur l'analyse de corpus, on peut s'attendre à une avancée importante sur la description du français parlé et simultanément sur la méthodologie de corpus entre les années soixante et les années deux mille dix. Si l'avancée a été majeure et déterminante pour les travaux sur la syntaxe du français, elle n'a apporté qu'une contribution très faible à la linguistique de corpus ou plus exactement à la linguistique *sur* corpus. La relation relativement distante entretenue entre les travaux du GARS et la sociolinguistique explique ce rendez-vous manqué.
Quatre disciplines vont avoir une incidence plus forte dans la même période sur les corpus de français parlé. Discipline compagne, la sociologie va opérer un lourd travail sur le recueil des données et sur la méthodologie d'entretien qui reste une part importante des corpus oraux. Parallèlement, la linguistique de l'interaction et plus particulièrement l'Analyse de Conversations va se développer très fortement et proposer une nouvelle approche du recueil de données « non provoquées par le chercheur ». Ensuite, le domaine de l'acquisition du langage fournira une méthodologie très rigoureuse de grandes bases de données partagées (volet français du programme CHILDES, notamment pour ce qui concerne l'adoption d'un format et d'un codage communs[1]) de corpus de productions d'enfants.
Enfin la recherche en technologies de la parole, de la reconnaissance à la synthèse en passant par la traduction repose sur le traitement de données orales massives.

La reprise du projet ESLO1 par l'équipe du CORAL (devenue LLL) en 2004 avec comme perspective de rendre disponible l'intégralité du corpus[2] et d'en constituer un nouveau devait

[1] MacWhinney B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ : Lawrence Erlbaum Associates.

[2] Un travail remarquable avait déjà été réalisé dans le cadre du projet ELILAP-ELICOP : ELILAP 1980-83 puis LANCOM 1993-2001, voir Mertens (2002)

nécessairement tenir compte des avancées apportées par ces disciplines.

Un bref bilan de l'impact de celles-ci révèle la qualité du travail précurseur des auteurs d'ESLO1 et facilite la reprise du projet avec une forte continuité même si plusieurs choix sont caractéristiques de l'évolution d'ESLO2.
Outre le soin apporté à la technique de conduite d'entretiens, les principales évolutions concernent l'intérêt accru pour assurer une représentation de l'hétérogénéité du panel de locuteur et des situations enregistrées (cf chapitre sur l'architecture du corpus en infra) et pour la description des langues en contact avec le français.

2.4 Conserver et diffuser le français ordinaire

Le bouleversement le plus fort concerne un élément peu fréquent jusqu'à très récemment dans les projets sur les corpus de français parlé : celui de la conservation et de la diffusion.
Pourtant sur ce point aussi ESLO1 était totalement précurseur.
Alors que dix ans auparavant les responsables du *Français fondamental* effaçaient les enregistrements réalisés dans le cadre de ce projet d'ampleur internationale (Abouda & Baude 2006), les auteurs d'ESLO1 décidaient d'apporter un soin particulier au catalogage de leurs enregistrements afin d'en assurer la meilleure diffusion. Ainsi un des six objectifs d'ESLO1 était de :

> *préparer et publier un catalogue descriptif et analytique des documents sonores et écrits, afin de les rendre disponibles aux chercheurs, notamment dans les domaines de la linguistique, de la sociologie et de la pédagogie des langues* (Lonergan, 1974 :2).

Cette volonté affichée dès l'origine du projet aura une forte incidence sur son développement. Elle porte la marque d'une relation particulière aux données et au rôle de leur exploitation partagée dans la constitution d'un savoir collectif. C'est également une reconnaissance de la légitimité de la langue

parlée comme objet scientifique et patrimonial. L'ESLO deviendra alors une référence sous le nom du *Corpus d'Orléans* et voyagera de la France à l'Angleterre, des Pays-Bas à la Belgique au gré des nombreux travaux de chercheurs dans une discipline en plein développement.

3. Le corpus des ESLO

3.1 Un très grand corpus

Le corpus des ESLO[3] a comme objectif d'être un très grand corpus de français parlé constitué de plusieurs centaines d'heures d'enregistrements afin d'atteindre une masse de 10 millions de mots.

Il est composé du corpus ESLO1, qui est un corpus clos, réalisé entre 1968 et 1971 et qui comprend 470 enregistrements d'une durée totale de 318 heures ce qui représenterait, selon l'estimation de l'époque, 4.5 millions de mots[4].

Le corpus ESLO2, en cours de réalisation, affiche un objectif de plus de six millions de mots pour 450 heures d'enregistrements.

Réunis dans une même base de données comprenant les enregistrements, leurs transcriptions orthographiques et les métadonnées décrivant les documents, le contexte d'enregistrement et les locuteurs, le corpus des ESLO est actuellement le plus grand corpus de français parlé disponible pour la recherche en linguistique.

L'objectif du projet n'est pas de produire un corpus représentatif, mais d'offrir un réservoir de corpus conçu dans un souci de représentativité des pratiques linguistiques d'une communauté d'auditeurs dans une ville donnée à des moments distincts. La sélection d'un sous-corpus d'études à partir de ces données reste à la charge du chercheur dans une démarche où la sélection des données est une étape fondamentale de l'analyse.

[3] Cf. Baude & Dugua 2011

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

Il revient alors aux auteurs des ESLO de rendre disponibles les données tout en les situant à la fois dans le cadre de leur contexte de production par les locuteurs et de celui de production par l'équipe scientifique y compris dans ses aspects et contraintes technologiques.
Il ne s'agit donc pas de produire un corpus de masse de données sans en préciser l'architecture et les cadres théoriques qui la conditionnent.

3.2 Architecture du corpus

La composition du corpus a subi une évolution sensible entre ESLO1 et ESLO2.
Comme nous l'avons indiqué, le corpus ESLO1 correspond déjà à une prise en charge des variations linguistiques selon différents axes. Cette recherche de la variation s'est concrétisée par une architecture qui en donnant une place centrale aux entretiens en face à face a néanmoins intégré sept autres modules dédiés à la diversité des situations de production de discours :
- Interviews sur questionnaires (interviews en face-à-face sur des questionnaires standardisés, avec un échantillon statistique aléatoire choisi d'après la liste INSEE du recensement de la population 1968). 157 enregistrements, 182,5 heures.
- Opérations sur le vif : contacts (prises de contact, reprises de contact, ouverture et clôture des entretiens enregistrés à l'insu du témoin). 55 enregistrements, 12,5 heures.
- Opérations sur le vif : témoins en situations sociales ou professionnelles (enregistrements de témoins INSEE dans des situations sociales ou professionnelles, faits en l'absence des chercheurs). 16 enregistrements, 14,5 heures.
- Communications téléphoniques. 50 enregistrements, 2,15 heures.
- Interviews sur mesure (entretiens avec des individus choisis selon leur rôle dans la "microsociété" orléanaise). 45 enregistrements, 48,33 heures.

- Conférences-débats (conférences-débats ou discussions à plusieurs participants, les dernières comportant souvent des témoins INSEE). 26 enregistrements, 34,15 heures.
- Enregistrements divers (enregistrements divers comportant des témoins inconnus, visite d'atelier, marchés, magasins, etc.). 84 enregistrements, 14,33 heures.
- CMPP (interviews au Centre Medico-Psychopédagogique, parents d'élèves et assistante sociale). 37 enregistrements, 10 heures.

L'ensemble de ces modules sont décrits dans le catalogue original (Lonergan, 1974 : 1) et présentés sur le site de diffusion du corpus ESLO[5].

L'architecture va considérablement évoluer dans le cadre du corpus ESLO2[6] afin de prendre en compte l'avancée méthodologique et théorique réalisée entre 1968 et 2008. D'une part l'évolution technologique a une forte incidence sur la collecte des corpus oraux. Si les auteurs d'ESLO1 se félicitaient de disposer de matériel d'enregistrement peu volumineux (de la taille d'une petite valise), et léger (à peine 7 kilos), l'équipe d'ESLO2 dispose d'un matériel numérique offrant les possibilités d'équiper des locuteurs de micro cravates HF pour une qualité d'enregistrement de tout premier ordre. Ainsi pour l'un des modules qui consiste à enregistrer l'intégralité de ce qu'une personne entend pendant 24 heures, les locuteurs sont équipés d'un micro les accompagnant dans toutes les activités de la vie quotidienne, de la toilette à la soirée entre amis en passant par l'activité professionnelle et les conversations familiales.

Cette évolution technologique s'accompagne d'un engouement fort pour la captation d'enregistrements les plus diversifiés dans des situations non provoquées par le chercheur selon les objectifs de l'Analyse de conversations.

[5] http://eslo.huma-num.fr/

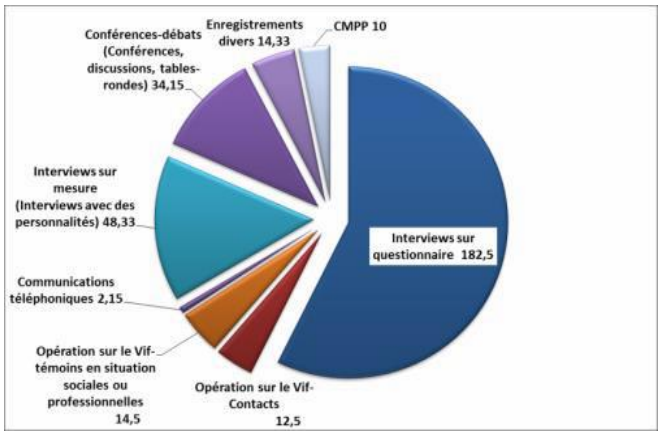
[6] http://eslo.huma-num.fr/index.php/pagecorpus/pagepresentationcorpus

L'objectif de dresser un portrait sonore ne peut donc se résumer à la collecte d'entretiens selon un échantillonnage sociologique. Il convient également d'élaborer une architecture de corpus qui permet de rendre compte de la diversité des situations de production et d'audition. Force est de constater qu'ESLO1 était balbutiant sur cet aspect. Si les entretiens ont été réalisés avec beaucoup de rigueur, les autres types d'enregistrements sont très souvent de très mauvaise qualité et correspondent à des objectifs peu maitrisés. La tentative d'enregistrer la même personne dans diverses situations s'est réduite à de simples tests sur quelques locuteurs. ESLO2 a donc comme ambition de présenter une forte évolution de la méthodologie de collecte de situations variées et représentatives des pratiques d'une communauté.

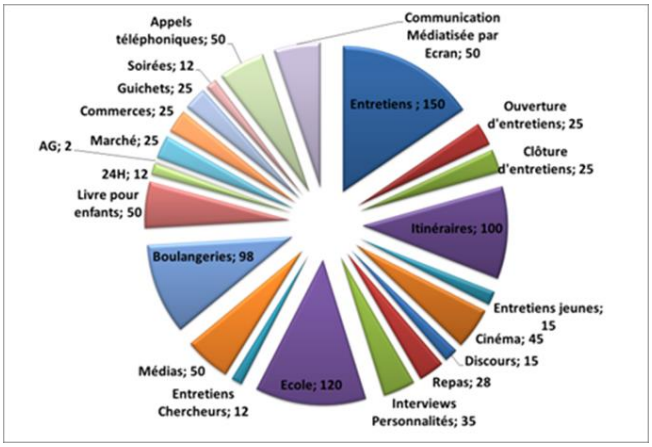
C'est toute l'architecture du corpus qui doit être modifiée afin de prendre en compte une grande diversité de situations de productions linguistiques tout en les situant au sein d'un marché linguistique plus général.

Le premier effet de ce changement est de pondérer la place des entretiens par rapport à d'autres types d'enregistrements. Les graphiques suivants qui expriment en nombre d'heures et en pourcentage la place de chacun des modules pour les deux corpus, rendent compte de ce changement.

ESLO1 :



ESLO2 :



3.3 Catégorisation des modules

L'architecture d'un corpus ne peut se résumer au pourcentage des genres, styles ou situations représentées. Elle nécessite également une réflexion sur la pertinence de ces catégories au sein d'une structure globale.

Ainsi, assurer la collecte de la diversité des pratiques linguistiques répond à un objectif d'enquête sociolinguistique et de description linguistique. Le conditionnement en corpus numérique du résultat de cette collecte nécessite un travail de catégorisation des modules constituant l'architecture du corpus. Cette catégorisation se doit d'être explicite et disponible à des fins de traitement des données. La classification habituelle dans les corpus de français parlé repose sur une opposition simpliste entre discours public et discours privé décrivant le niveau de formalité des énoncés.

Ainsi, *Le Corpus de référence du français parlé*, réalisé par Claire Blanche Benveniste et l'équipe DELIC à partir de 1998, repose sur une structure en trois modules : parole privée, parole professionnelle et parole publique. Cette distinction est assez rudimentaire si on se réfère aux travaux de l'analyse de

conversations ou même à la description des registres de langue (Koch & Oesterreicher 2001).

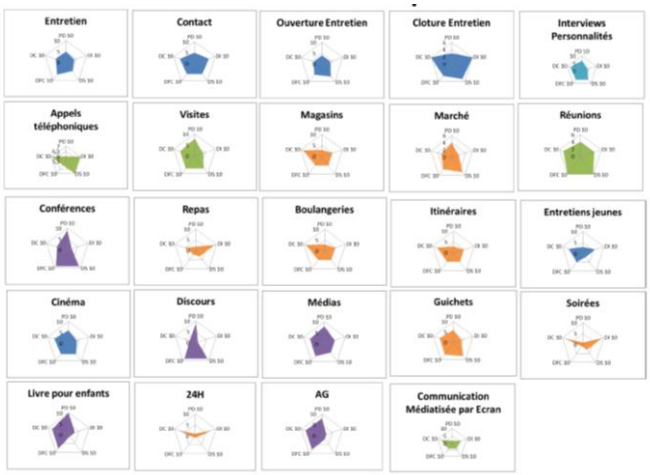
Le corpus ESLO2 est l'occasion de tenter une description des registres, styles ou types de situations en partant des caractéristiques apriori et, apostériori, des différents modules.

Chaque module est décrit a priori, c'est-à-dire avant la collecte et non sur la base d'une analyse du contenu, selon les critères suivants :

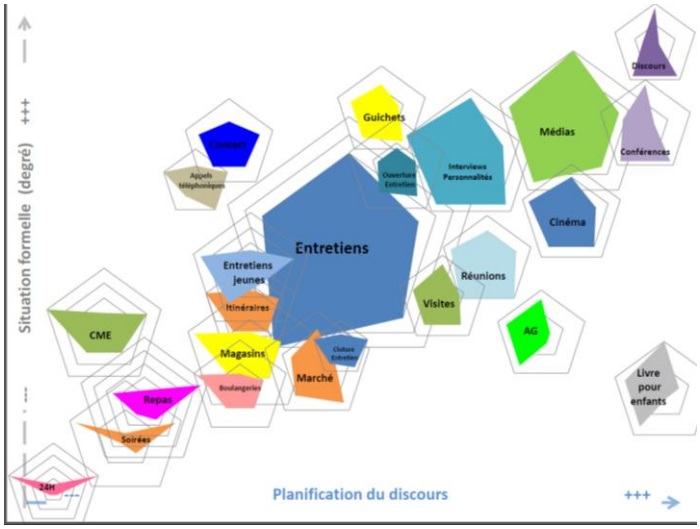
- Degré de planification du discours (en opposant le registre « spontané » de la conversation ordinaire à celui de conférences ou le discours est écrit),
- Degré d'interactivité (du monologue au dialogue et autres conversations relevant d'un travail conséquent d'interaction),
- Degré de distance sociale entre les interactants (à partir des critères traditionnels de la sociologie : âge, sexe, niveau d'études, profession),
- Degré de convergence (de la polémique au consensus),
- Degré de formalité du cadre (au sens de Goffman, chaque situation pouvant se définir selon un cadre social impliquant des statuts, rôles et comportements langagiers).

Chacun de ces critères est évalué sur une échelle de 0 à 10 et le module peut être visualisé selon la forme obtenue par un graphique en radar :

Les différents modules constitutifs de l'arcitecture ESLO2 :



Cette démarche permet de décrire l'architecture du corpus en raffinant une prise en compte des axes traditionnels qui situent un contexte de production de discours selon le degré de formalisme de la situation sociale d'une part et le degré de planification de l'énoncé d'autre part.



Cette représentation de l'architecture du corpus répond à deux objectifs. Premièrement, il s'agit de définir avec précisions les différents modules qui composent le corpus complet en situant les situations enregistrées selon les critères de la sociologie et de la pragmatique. Cela répond à une conception des pratiques linguistiques comme relevant systématiquement d'un *contexte*, qui n'est autre qu'un marché linguistique au sein duquel les locuteurs mobilisent des comportements langagiers dans un but d'interaction.
Deuxièmement, l'évaluation des modules selon différents critères permet un travail réflexif sur une définition apriori et un constat apostériori à partir des données précises de la situation enregistrée. Ainsi, si le module entretien répond globalement à une définition selon les critères présentés, celle-ci va être pondérée pour chaque entretien. L'évaluation de la distance sociale et du degré d'interactivité peuvent par exemple être très différents d'un entretien à l'autre et déboucher sur une représentation proche d'une conversation ordinaire dans un cas ou d'un discours public ou médiatique dans un autre.
In fine, cette réflexion sur l'architecture du corpus permet de concevoir ESLO2 comme un corpus ouvert sans pour autant le réduire à un empilement, opportuniste et sans fin, d'enregistrements variés.

3.4 État du corpus

L'ensemble des enregistrements est maintenant numérique. L'intégralité des enregistrements ESLO1 a été numérisée dans le cadre du dépôt du fonds à la Bibliothèque Nationale de France. ESLO2 est nativement collecté en numérique à l'aide de différents matériels selon les contraintes des modules[7]. Si ESLO1 est un corpus clos, la collecte d'ESLO2 continue à la date de la rédaction de cet article.

Tous les enregistrements sont catalogués et indexés (cf. chapitre suivant) et la transcription de l'intégralité des corpus est en cours.

[7] Principalement : enregistreurs Marantz PMD 661 MKII + microcravates AKG C417L, TASCAM DR100, Edirol R09 : http://eslo.huma-num.fr/index.php/pagemethodologie?id=70

Les opérations de formatage, catalogage et transcription sont excessivement lourdes ce qui explique le peu de corpus d'envergure disponibles. Face à cette difficulté, les chercheurs se replient souvent vers un usage du corpus restreint à leur recherche. La particularité forte du projet des ESLO est au contraire de maintenir un objectif scientifique clairement identifié tout en attribuant au corpus une valeur patrimoniale et scientifique qui dépasse le cadre du projet initial. Il en résulte un vaste chantier de traitement du corpus qui sera détaillé dans la dernière partie de cet article. Nous pouvons néanmoins faire état de l'avancement de ces opérations. Ainsi, au premier mai 2015 le corpus des ESLO est composé de :

	Enregistrements		**Transcrits**	
	Nbre	Heures	Nbre	Heures
ESLO1	468	318	336	274
ESLO2	590	266	583	259
TOTAL	**1058**	**584**	**919**	**533**

4. Un corpus pour les *Humanités numériques*

4.1 Le temps des humanités numériques

Le projet de diffusion des ESLO au début des années 2000 est contemporain de la mutation des sciences humaines et sociales dans ce qu'on appelle dorénavant le tournant des *Digitals Humanities* ou *Humanités numériques* voire *Humanités digitales* (Le Deuff, 2014)[8]. Les discussions sur ce que sont les *humanités numériques* sont très vives et la définition reste très ouverte. Il ne s'agit pas de rentrer ici dans une vaste discussion sur la pertinence d'une approche en terme de naissance d'une discipline, d'une trans-discipline ou d'une appropriation d'outils numériques par des disciplines traditionnelles, nous nous contenterons de constater que la linguistique est en

[8] Le Deuff O. dir. (2014). *Le temps des humanités digitales,* la mutation des sciences humaines et sociales.

première ligne d'un questionnement sur les conditions de constitution, de diffusion et de partage d'un savoir transformé par le croisement de l'informatique, du numérique et des arts et lettres au sein des sciences humaines et sociales. Ces grands principes ont été définis dans le *Manifeste des Digital humanities*[9].

D'une manière plus concrète encore nous présentons ici les principales caractéristiques qui inscrivent le projet des ESLO dans cette approche des corpus en sciences humaines et sociales. Le soin apporté à la diffusion d'ESLO1 en 1974 en réalisant un « *catalogue descriptif et analytique des documents sonores et écrits, afin de les rendre disponibles aux chercheurs (Lonergan 1974 : 2)* » peut être interprété comme la première pierre posée dans l'édifice d'un corpus qui dépasse les enjeux de l'étude des auteurs. La seconde pierre viendra de l'équipe de Piet Mertens et du projet ELICOP quelque trente ans plus tard en rendant accessible une partie du corpus après un lourd travail de normalisation des conventions de transcription et même d'annotations morphosyntaxiques contenues dans des balises au format SGML. Ce travail s'appuie sur les perspectives dressées par la linguistique de corpus telle qu'elle est définie par Habert, Nazarenko et Salemn en 1997, mais n'est pas encore directement orienté vers un traitement d'ensemble.

C'est à partir de 2004 avec la numérisation d'ESLO1 et le souhait de rendre le corpus intégralement disponible pour des usages scientifiques mais aussi culturels que l'édifice s'ancrera définitivement dans les humanités numériques.

4.2 L'interopérabilité et l'archivage

La question de la réutilisation d'un corpus n'est pas anodine et ne va pas de soi. Il ne s'agit pas ici d'affirmer que toute recherche linguistique doit s'appuyer sur un corpus et que tout corpus peut être réutilisé pour d'autres recherches. Rien n'est moins sûr mais dans le cas des ESLO c'est un parti pris affirmé par les différents auteurs du projet. Le périmètre du projet est de

[9] http://tcp.hypotheses.org/318

fait vaste, il s'agit de produire un portrait sonore d'une ville en faisant l'hypothèse que le corpus produit peut être utile à diverses recherches en linguistique, sociologie, histoire, didactique et acquiert ainsi une dimension patrimoniale qui a également pour effet de légitimer le français tel qu'il est parlé dans sa très grande diversité.

L'objectif affirmé est donc de disposer de données répondant à un critère d'interopérabilité. Celui-ci se concrétise à différents niveaux.

Premièrement, les enregistrements sont conservés dans un format numérique selon les recommandations d'une structure internationale, l'*International Association of Sound and Audiovisual Archive*[10].

Deuxièmement, les documents sont systématiquement accompagnés de métadonnées descriptives. Le choix retenu est celui du format *DUBLIN-CORE Open Language Archives Community*[11]. Il s'agit d'un choix minimal qui a été repris dans le cas de diffusions liées à d'autres objectifs. Ainsi le format CMDI[12] est celui utilisé dans la perspective européenne CLARIN, le format EAD[13] par la BnF pour l'intégration à son catalogue *Archives et manuscrits*, et l'EDM dans le cadre de la bibliothèque européenne *Europeana*[14].

Troisièmement, les enregistrements sont transcrits et synchronisés avec le signal sonore selon des conventions minimales[15] répondant à un format interopérable. Le format choisi est un format XML qui est ensuite repris pour un enrichissement en TEI (TEIML[16]). Les transcriptions sont segmentées en unités prosodiquement, syntaxiquement et

[10] http://www.iasa-web.org/ : Wave, stéréo, 16 bits, 44100 Hz.

[11] http://www.language-archives.org/OLAC/metadata.html

[12] http://www.clarin.eu/content/component-metadata

[13] http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_ead.html

[14] http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

[15] http://eslo.huma-num.fr/index.php/pagemethodologie?id=71

[16] Norme ISO/CD 24624 en cours d'élaboration

sémantiquement cohérentes afin d'assurer une synchronisation à l'aide de jalons temporels fréquents. La transcription proposée repose sur des conventions minimales. A ce stade, il s'agit de répondre à un simple objectif de navigation dans le corpus. Pour toute analyse ultérieure une reprise de la transcription avec des conventions répondant aux cadres théoriques du chercheur est indispensable.

L'ensemble de ces choix permet l'utilisation d'un service d'archivage. Expérimenté dans le cadre du projet pilote sur l'archivage de l'oral par le TGE ADONIS puis poursuivi par la TGIR HUM-NUM, les données (enregistrements, transcriptions et métadonnées) sont confiées à la plateforme Cocoon[17] qui en assure le stockage sécurisé sur la grille Huma-Num hébergée au centre de calcul de l'IN2P3. Pendant cette phase, Cocoon assure des services de contrôle de la qualité des données puis verse les données au Centre Informatique National de l'Enseignement Supérieur pour une conservation intermédiaire avant de rejoindre les Archives Nationales pour un archivage définitif. Parallèlement, les bandes magnétiques originales ont été confiées au service sonore du département de l'audio-visuel de la BnF.
Les opérations d'archivage sont également l'occasion d'attribuer un identifiant unique et pérenne à tous les documents constitutifs du corpus.

4.3 Les aspects juridiques

La diffusion du corpus est bien évidemment liée à des aspects juridiques. Sur ce point, le projet a bénéficié du travail diffusé par le *Guide des bonnes pratiques 2006*[18].
Le choix de l'équipe a été d'apporter beaucoup d'attention à une démarche éthique en recueillant le consentement éclairé de

[17] http://cocoon.huma-num.fr/exist/crdo/

[18] Baude et al., (2006).

toutes les personnes enregistrées[19]. Les enregistrements et les transcriptions sont également anonymisés et les données personnelles conservées dans une base de données séparée.
Les données sont diffusées sous licence creatives commons[20] (BY NC SA : Attribution, pas d'utilisation commerciale et partage dans les mêmes conditions) : le titulaire des droits autorise l'exploitation de l'œuvre originale à des fins non commerciales, ainsi que la création d'œuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'œuvre originale.

4.4 Le signalement et la diffusion

La conservation des données étant assurée à différents niveaux (stockage sécurisé, conservation intermédiaire et archivage pérenne) et les aspects juridiques ouverts à une large diffusion, il faut en assurer l'accès pour différents usages.
Sur ce point, le soin apporté à l'interopérabilité devient crucial.
Les données ESLO sont accessibles sur un site dédié au projet[21], géré par l'équipe du Laboratoire Ligérien de Linguistique et hébergé sur la grille Huma-Num.
Le site, réalisé à l'aide du CMS Joomla et intégrant une application, a été conçu en trois parties :
- Une interface « back office » qui permet la gestion du corpus. Cette interface permet, à l'aide de formulaires, de renseigner les métadonnées et dispose de fonctionnalités pour attribuer aléatoirement les identifiants anonymes, transférer les fichiers sonores et les transcriptions sur la plateforme Cocoon et pour accéder à une base de données mysql qui contient les transcriptions et les métadonnées.
- Une interface d'accès aux corpus avec des outils spécifiques. L'accès aux corpus se fait par une recherche des documents dans leur intégralité sous la forme d'un catalogue ou par la recherche d'une chaine de caractères au sein des transcriptions.

[19] http://eslo.huma-num.fr/index.php/pagemethodologie?id=69

[20] http://creativecommons.fr/licences/les-6-licences/

[21] http://eslo.huma-num.fr/

Un outil de requête permet de croiser les critères de recherche sur les transcriptions avec les informations sur les documents et les locuteurs.
Un second outil offre la possibilité d'écouter l'enregistrement synchronisé sur le signal.
Enfin, l'ensemble des documents sont téléchargeables directement soit pour tout utilisateur du site soit pour un utilisateur ayant signé une convention lorsqu'il y a des restrictions juridiques.
- La dernière fonctionnalité du site est d'offrir un contenu éditorial principalement orienté vers les documents méthodologiques : conventions et guides de transcriptions, documents techniques et juridiques, documents scientifiques.

Cette diffusion du corpus par un site spécifique répond principalement aux objectifs du Laboratoire Ligérien de Linguistique. La gestion des données selon des bonnes pratiques d'interopérabilité et d'archivage permet un signalement et une diffusion beaucoup plus large.

Ainsi, la plateforme Cocoon propose un entrepôt exposant les métadonnées en Open Archive Initiative. Le corpus des ESLO est donc signalé par tout instrument reposant sur un moissonnage en OAI. C'est notamment le cas de la plateforme ISIDORE[22] qui permet la recherche et l'accès aux données numériques en sciences humaines et sociales. Au premier mai 2015, une recherche sur ESLO dans le moteur d'ISIDORE apporte 2001 réponses, soit l'ensemble des documents disponibles à ce moment-là dans la collection ESLO de l'entrepôt Cocoon.

Comme ESLO existe également sous la forme de bandes magnétiques originales conservées et décrites par la BnF, le corpus est également signalé dans ses catalogues.

[22] http://www.rechercheisidore.fr/

Enfin le corpus des ESLO a été naturellement intégré à l'EQUIPEX ORTOLANG[23] dont l'objectif est de gérer une « *infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés* ».

4.5 Le web de données

Le travail sur la structuration des données et des métadonnées et la gestion de la diffusion du corpus des ESLO permet un travail exploratoire dans le cadre du web de données (ou web sémantique). Cette étape concrétise la volonté de construire un corpus réutilisable pour une grande variété d'usages. Le web de données vise à publier des données structurées sur le Web, afin de les relier entre elles et donc d'enrichir un réseau d'informations. Elle nécessite l'utilisation, dans un format spécifique, de vocabulaires, référentiels et ontologies facilitant le liage des données.

Nous pouvons citer quelques exemples d'expérimentations en cours auxquelles participe ESLO :

- la plateforme ISIDORE qui repose sur les principes du Web de données,
- data.bnf.fr, le projet qui donne accès aux données contenues dans ses catalogues et dans Gallica,
- le programme *Sémantisation du Corpus de la parole* du Ministère de la Culture,
- le projet « Cabinet de curiosités des langues de France » réalisé dans le cadre de l'appel à propositions « services culturels innovants du Ministère de la culture ».

Ces différents projets sont trop récents pour en tirer un premier bilan. Un seul exemple peut néanmoins démontrer l'intérêt de rendre un corpus disponible selon les pratiques en vigueur dans le domaine du web de données. Une recherche sur le terme « abattoirs » permet, par l'outil data.bnf.fr de signaler, d'écouter et de télécharger l'enregistrement d'ESLO consacré à l'entretien d'un boucher d'Orléans, et la même requête sur

[23] https://www.ortolang.fr/

ISIDORE permet de trouver une correspondance entre cet enregistrement et un entretien sur le même thème réalisé par des sociologues à Toulouse dans les années 1960.

Conclusion

Le corpus des ESLO a été réalisé par des linguistes et il a donné lieu à de très nombreux travaux en linguistique. Après les différents travaux en phonologie, syntaxe, prosodie, lexique et autres domaines engendrés par ESLO1, l'équipe d'ESLO2 réalise différentes études directement issues d'une analyse du corpus ou fondées sur une comparaison avec d'autres corpus[24].
A partir d'ESLO1 une méthode d'apprentissage des langues particulièrement innovante[25] a été réalisée et des travaux sont en cours de réflexion dans le cadre d'un usage didactique du corpus ESLO2.
On peut donc considérer que l'objectif d'obtenir un portrait sonore d'une communauté d'auditeurs d'une même ville est une source importante d'études linguistiques et d'applications liées.
Il convient néanmoins d'être prudent, ce portrait sonore ne peut se résumer à des enregistrements divers et variés sans un cadre théorique qui fait de la linguistique de corpus une discipline qui doit entendre autant si ce n'est plus, la sociolinguistique que la linguistique outillée par l'informatique.
Le tournant des *humanités numériques* est l'occasion de repenser cette définition de la linguistique sur corpus afin de définir une véritable science des données linguistiques. Face à ce défi, le linguiste doit maitriser l'ensemble de la chaine qui le conduit à travailler, exploiter et diffuser ces données collectées qui ne lui sont jamais « données ». Il est aussi important qu'il prenne conscience que cette science relève d'un domaine au sein duquel il n'est pas le seul acteur.

[24] Comme par exemple les travaux sur la liaison dans ESLO, PFC et d'autres corpus (Dugua et Baude, à paraitre).

[25] Biggs P. & Dalwood M. 1976, *Les Orléanais ont la parole.*

Références bibliographiques

Site ESLO : http://eslo.huma-num.fr

Abouda L. & Baude O. (2009). « Du Français Fondamental aux ESLO », *in* Bruxelles, Mondada, Simon, Traverso (ed.) *Grand corpus de français parlé, Bilan historique et perspectives de recherche*. Cahiers de Linguistique Revue de sociolinguistique et de sociologie de la langue française 33/2, EME, Louvain, 131-146.

Abouda L. & Baude O. (2007). « Constituer et exploiter un grand corpus oral, choix et enjeux théoriques : le cas des ESLO », in actes du colloque Corpus en lettres et sciences sociales, *Des documents numériques à l'interprétation*, Colloque d'Albi Langages et Signification juin 2006, Presses universitaires de Toulouse: 161-168.

Baude O. & Bergounioux G. (à paraitre 2015). « L'ESLO : une enquête en son temps » *in Linguistique de corpus : une étude de cas La recette de l'omelette*, Champion.

Baude O. & Lacheret A. (à paraitre 2015). "The collection of data for the Rhapsodie Treebank: typological criteria and ethical issues" in A. Lacheret, S. Kahane & P. Pietrandrea (ed.) *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French, col Studies in Corpus Linguistics*. Amsterdam, Benjamins.

Baude O. & Dugua C. (2011). « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? » Corpus 10, *Varia*, 99-118.

Baude O. coord. (2006). *Corpus oraux, guide des bonnes pratiques*, Paris et Orléans, Editions du CNRS et Presses Universitaires d'Orléans.

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.

Biggs P. & Dalwood M. (1976). *Les Orléanais ont la parole : Teaching Guide and Tapescript*, Londres, Longman (Livre du maître).

Biggs P. & Dalwood M. (1976). *Les Orléanais ont la parole*, Londres, Longman (Livre de l'élève).

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

Blanche-Benveniste C. et alii (1990). *Français parlé. Etudes grammaticales*. Paris : CNRS.

Bourdieu P., Chamboredon 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.

Bourdieu P. (1984). « Le marché linguistique », *Questions de sociologie*, éditions de Minuit, Paris.

Equipe DELIC (2004). Recherches sur le français parlé n° 18, Publications de l'université de Provence, 265 p.

Encrevé P. (1976). « Présentation », *in* W. Labov, *Sociolinguistique*, éditions de Minuit, Paris.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C. & Tellier I. (2012). « Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012 », in *Ressources linguistiques libres, TAL*. Volume 52 – n° 3/2011, 17-46.

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

Jacobson M. & Baude O. (2012). « Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France », in *Ressources linguistiques libres, TAL*. Volume 52 – n° 3/2011, 47-69.

Koch P. & Oesterreicher W. (2001). « Langage oral et langage écrit », *Lexicon der Romanistischen Linguistik*, tome 1-2, Tübingen, Max Niemeyer, 584-627.

Laks B. (2013). « Why is there variation instead of nothing », *Language Sciences* 39: 31-53

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

Le Deuff O. dir (2014). *Le temps des humanités digitales*, Limoges : FYP éditions.

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.

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.