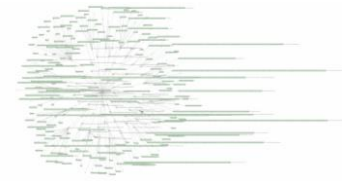
Observatologie : Vers une science de l'adéquation observationnelle en linguistique

Olivier Baude

université
Paris Ouest
Nanterre La Défense

Observatologie :

Vers une science de l'adéquation observationnelle en linguistique



Volume de synthèse de l'Habilitation à diriger des recherches

Présentée par

Olivier Baude

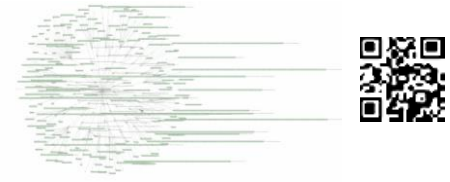
Le 27 novembre 2015

Jury :

M. Bernard Laks, Professeur à l'université Paris Ouest (référent)
M. Gabriel Bergounioux, Professeur à l'université d'Orléans
Mme Isabelle de Lamberterie, Directrice de recherche au CNRS
M. Loïc Depecker, Professeur à l'université Sorbonne Nouvelle Paris 3
M. Pierre Encrevé, Directeur d'études à l'EHESS
Mme Annette Gerstenberg, Professeure à l'Université Libre de Berlin (FUB)
M. Jean-Luc Minel, Professeur à l'université Paris Ouest

Observatologie :

Vers une science de l'adéquation observationnelle en linguistique



Volume de synthèse de l'Habilitation à diriger des recherches

Olivier Baude

Cette synthèse présente un parcours de 15 ans de travail dont la caractéristique première est de se situer à un niveau collectif. Il serait donc impossible de citer toutes les personnes qui ont eu un rôle essentiel dans celui-ci. Je sais néanmoins ce que je dois, en tout premier lieu à Pierre Encrevé et à ceux qui ont été des repères fondamentaux : Bernard Lak et Gabriel Bergounioux mais aussi Michel de Fornel Benoit Habert, Françoise Gadet, Isabelle de Lamberterie, Françoise Genova, Michel Alessio, Lorenza Mondada et Pascal Cordereix. Je dois énormément à Michel Jacobson et encore plus à Céline Dugua, sans eux les projets exposés dans cette synthèse n'auraient jamais vu le jour.
Le niveau collectif des projets commence avant…Laurence, mes enfants et petits-enfants sont au cœur de ce travail.

Nota : lecture du document

Cette synthèse de travaux, rédigée dans le cadre d'une Habilitation à Diriger des Recherches, présente trois spécificités de forme qui sont des éléments assumés de la narration d'une activité scientifique (argumentation détaillée dans l'introduction) :

1. La cohérence de 15 ans de recherche est présentée à partir de cartes heuristiques qui en permettent une lecture thématique. Le lecteur peut ainsi avoir une vue d'ensemble et choisir différents parcours de lecture.

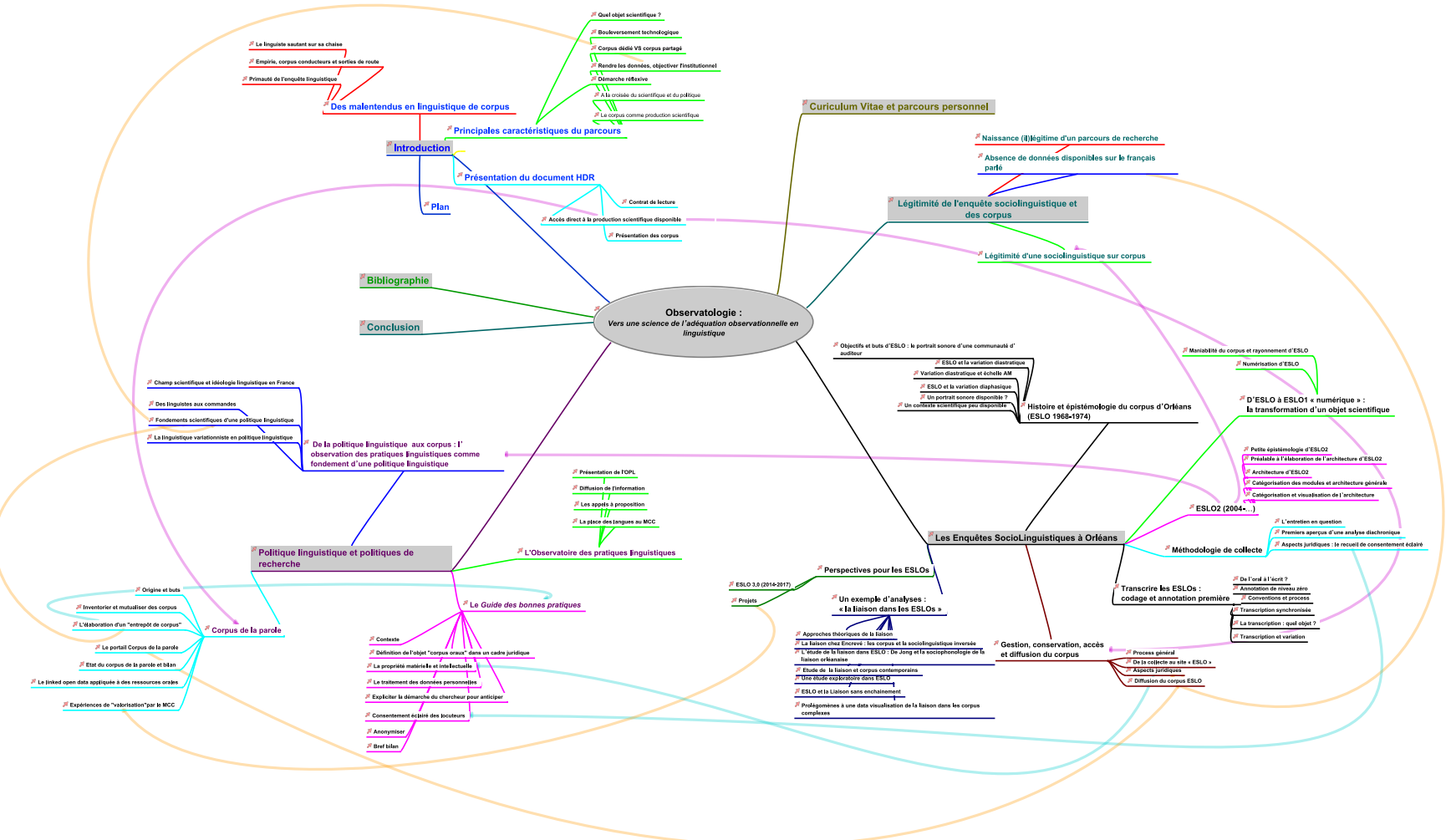
Un clic sur les flèches rouges  permet de faire paraitre le texte correspondant. L'icône *[retour]* permet de revenir à la carte.

Toutefois une version linéaire, classique, est également présentée sous la forme d'un PDF. Dans ce cas un certain nombre de liens internes (repérés ainsi) permettent néanmoins de naviguer dans l'ensemble du document pour suivre une thématique.

2. L'accent a été mis sur l'activité scientifique disponible de manière pérenne sur le Web. Ainsi, les liens bleus donnent accès aux articles, ouvrages, travaux et documents en ligne.

3. Les corpus sont au centre de cette activité scientifique. Ils sont donc intégrés au volume « Travaux » et une présentation détaillée permet d'évaluer le volume de travail que représente cette activité scientifique qui commence tout juste à être reconnue en tant que telle. Dans le cadre de cette Habilitation, le volume de travail consacré aux corpus est évalué à plusieurs dizaines d'articles. Ne sont présentés ici que les éléments des corpus disponibles, un permalien (permet de s'y référer systématiquement).

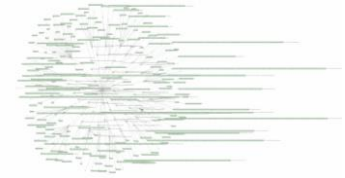
Carte heuristique


Observatologie :
Vers une science de l'adéquation observationnelle en linguistique
Introduction
Des malentendus en linguistique de corpus
Le linguiste sautant sur sa chaise
Empirie, corpus conducteurs et sorties de route
Primauté de l'enquête linguistique
Principales caractéristiques du parcours
Quel objet scientifique ?
Bouleversement technologique
Corpus dédié VS corpus partagé
Rendre les données, objectiver l'institutionnel
Démarche réflexive
A la croisée du scientifique et du politique
Le corpus comme production scientifique
Présentation du document HDR
Contrat de lecture
Accès direct à la production scientifique disponible
Présentation des corpus
Plan
Bibliographie
Conclusion
Curiculum Vitae et parcours personnel
Légitimité de l'enquête sociolinguistique et des corpus
Naissance (il)légitime d'un parcours de recherche
Absence de données disponibles sur le français parlé
Légitimité d'une sociolinguistique sur corpus
Histoire et épistémologie du corpus d'Orléans (ESLO 1968-1974)
Objectifs et buts d'ESLO : le portrait sonore d'une communauté d'auditeur
ESLO et la variation diastratique
Variation diastratique et échelle AM
ESLO et la variation diaphasique
Un portrait sonore disponible ?
Un contexte scientifique peu disponible
Les Enquêtes SocioLinguistiques à Orléans
D'ESLO à ESLO1 « numérique » : la transformation d'un objet scientifique
Maniabilité du corpus et rayonnement d'ESLO
Numérisation d'ESLO
ESLO2 (2004-...)
Petite épistémologie d'ESLO2
Préalable à l'élaboration de l'architecture d'ESLO2
Architecture d'ESLO2
Catégorisation des modules et architecture générale
Catégorisation et visualisation de l'architecture
Méthodologie de collecte
L'entretien en question
Premiers aperçus d'une analyse diachronique
Aspects juridiques : le recueil de consentement éclairé
Transcrire les ESLOs : codage et annotation première
De l'oral à l'écrit ?
Annotation de niveau zéro
Conventions et process
Transcription synchronisée
La transcription : quel objet ?
Transcription et variation
Gestion, conservation, accès et diffusion du corpus
Process général
De la collecte au site « ESLO »
Aspects juridiques
Diffusion du corpus ESLO
Perspectives pour les ESLOs
ESLO 3.0 (2014-2017)
Projets
Un exemple d'analyses : « la liaison dans les ESLOs »
Approches théoriques de la liaison
La liaison chez Encrevé : les corpus et la sociolinguistique inversée
L'étude de la liaison dans ESLO : De Jong et la sociophonologie de la liaison orléanaise
Etude de la liaison et corpus contemporains
Une étude exploratoire dans ESLO
ESLO et la Liaison sans enchainement
Prolégomènes à une data visualisation de la liaison dans les corpus complexes
Politique linguistique et politiques de recherche
De la politique linguistique aux corpus : l'observation des pratiques linguistiques comme fondement d'une politique linguistique
Champ scientifique et idéologie linguistique en France
Des linguistes aux commandes
Fondements scientifiques d'une politique linguistique
La linguistique variationniste en politique linguistique
L'Observatoire des pratiques linguistiques
Présentation de l'OPL
Diffusion de l'information
Les appels à proposition
La place des langues au MCC
Le Guide des bonnes pratiques
Contexte
Définition de l'objet "corpus oraux" dans un cadre juridique
La propriété matérielle et intellectuelle
Le traitement des données personnelles
Expliciter la démarche du chercheur pour anticiper
Consentement éclairé des locuteurs
Anonymiser
Bref bilan
Corpus de la parole
Origine et buts
Inventorier et mutualiser des corpus
L'élaboration d'un "entrepôt de corpus"
Le portail Corpus de la parole
Etat du corpus de la parole et bilan
Le linked open data appliqué à des ressources orales
Expériences de "valorisation"par le MCC

Olivier Baude

Observatologie :

Vers une science de l'adéquation observationnelle en linguistique



Résumé

Observatologie :
Vers une science de l'adéquation observationnelle en linguistique

Ce texte présente les éléments d'un parcours de chercheur vers une science de *l'observation linguistique*. Il s'agit de porter un regard réflexif et de situer la cohérence de différentes activités de recherche qui construisent, in fine, un chemin non linéaire à partir des questions des données de la recherche en linguistique vers la construction et la définition d'un objet scientifique.

Résolument ancré dans le champ d'une linguistique qui refuse d'établir une dichotomie entre la linguistique et la sociolinguistique parce qu'il part du principe que la langue est *par nature sociale*, ce travail est orienté vers la quête de l'adéquation observationnelle. En dépassant le cadre de l'enquête sociolinguistique d'un côté et de la linguistique de corpus de l'autre, il interroge ce que peut être une pratique et une théorie des données linguistiques et de leur condition de production

Le cheminement retracé dans ce document se situe au confluent de l'enquête linguistique et de la linguistique de corpus dans une période épistémologique qui correspond, depuis une dizaine d'années, à l'émergence du domaine des « *humanités numériques* ». Au cœur de cette approche il y a la part essentielle que prennent les données dont on ne peut séparer la collecte de l'exploitation, la méthodologie de la théorie, le terrain de l'analyse, la science de la politique. Leur convergence dessine les contours d'une véritable *science de l'observation* des données linguistiques.

Abstract

Observatology :
Towards a science of observational adequacy in linguistics

This text presents elements of a researcher's pathway towards a science of the linguistic observation. Our aim is to lead a reflection (on) and to situate the coherence of our different research activities that build, in fine, a non-linear way towards the construction and the definition of a scientific object, starting from the issue of research data in linguistics.

Firmly anchored in the field of linguistics that refuses the dichotomy between linguistics and sociolinguistics, and in line with the principle of the social nature of language, this work is oriented towards the quest of an observational consistency. Going beyond the framework of the sociolinguistic survey on one hand and of the corpus linguistics on the other hand, we question what can be a practice and a theory of linguistic data and the condition of their production.

Thus, the pathway presented in this text is intended to be at the confluence of linguistic survey and of corpus linguistics in an epistemological period that gave birth, ten years ago, to the field of "digital humanities". In the heart of this approach remains the essential share of the data in which we can't separate collection and exploitation, methodology and theory, field and analysis. All those elements should be gathered around a real science of the observation of linguistic data.

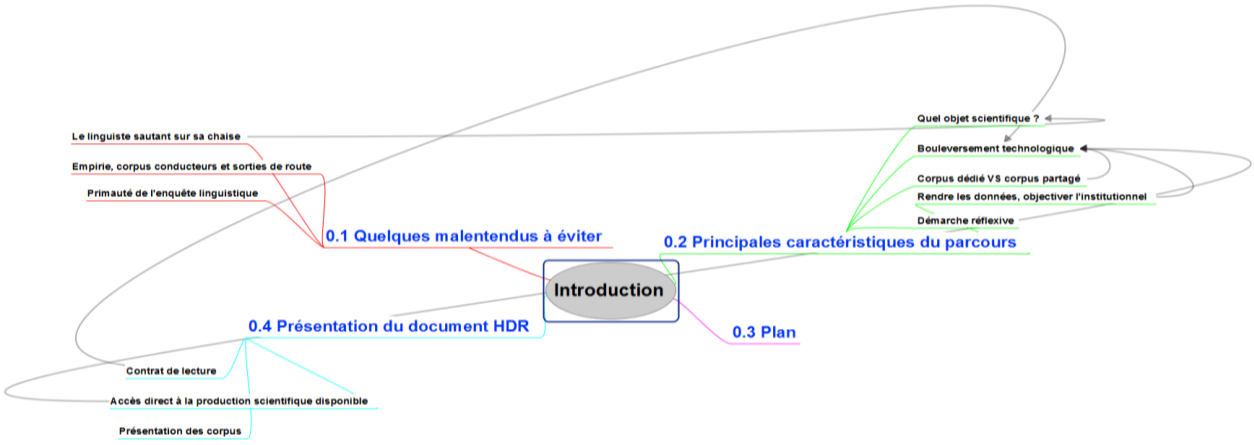
Table des matières

Contenu

Observatologie :

Vers une science de l'adéquation observationnelle en linguistique

0. Introduction *[retour]*



[1]Ce texte présente les éléments d'un parcours de chercheur vers une science de *l'observation linguistique*. Il s'agit de porter un regard réflexif et de situer la cohérence de différentes activités de recherche qui construisent, in fine, un chemin non linéaire à partir des questions des données de la recherche en linguistique vers la construction et la définition d'un objet scientifique.

Résolument ancré dans le champ d'une linguistique qui refuse une dichotomie entre linguistique et sociolinguistique, et donc qui part du principe que la langue est *par nature sociale* (Encrevé), ce travail est orienté vers la quête de l'adéquation observationnelle. Comme le souligne Encrevé :

> *« En 1962, au IX^e^ Congrès international des linguistes réunis à Cambridge, Noam Chomsky énonçait les trois niveaux d'adéquation que la linguistique devait atteindre : adéquation observationnelle, descriptive, explicative ; mais il délaissait le premier pour s'attaquer aux deux autres (…) Car la question des données va entrainer toutes*

[1] Ce texte est rédigé selon les recommandations orthographiques de 1990.

> *les autres. Partie structurée d'un tout qu'elle structure, la langue n'est en effet 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)*[2]

Dans les quarante années qui suivirent ce texte, la linguistique variationniste apporta de nouveaux cadres théoriques et méthodologiques. D'une manière étrange, c'est en parallèle que la linguistique « sur des données » se développa autour d'une conception particulière de la linguistique de corpus.

Le parcours présenté dans ce texte se veut au confluent de l'enquête linguistique et de la linguistique de corpus dans une période épistémologique qui voit naitre depuis une dizaine d'années le domaine des « *humanités numériques* ». Au cœur de cette approche reste la part essentielle des données dont on ne peut séparer la collecte de l'exploitation, la méthodologie de la théorie, le terrain de l'analyse. Tout ceci devant se regrouper autour d'une véritable *science de l'observation* des données linguistiques.

0.1 Des malentendus en linguistique de corpus [cadres corpus] [retour]

Nous pouvons écarter dès le début de ce travail certains malentendus :

Le linguiste sautant sur sa chaise : « Le corpus ! le corpus ! le corpus ! »

Il ne s'agit pas de concevoir exclusivement la linguistique comme ne pouvant exister qu'à partir d'une méthodologie de corpus. Le recours fréquent aux corpus comme simple réservoir d'exemples attestés ou comme argument d'autorité, souvent appuyé par l'argument quantitatif, est un risque fort de dérive.

Les études sur exemples, sur les représentations ou sur « la langue possible » sont souvent bien plus pertinentes qu'une linguistique de corpus qui ne s'interroge pas sur les données qu'elle traite.

Il existe néanmoins une conception de la linguistique qui postule l'adéquation observationnelle comme inhérente à la description à l'analyse linguistique. Il s'agit alors de construire une observation systématique des pratiques linguistiques dans leur contexte social avec la plus grande rigueur afin d'atteindre la nature sociale de la langue. Si l'on suit ce postulat il n'est pas possible de se contenter d'une linguistique de corpus non fondée sur une approche sociolinguistique. Il ne s'agit donc pas d'opposer « linguistique de corpus » à « linguistique sans corpus », mais plutôt « linguistique de l'observation des pratiques » à « linguistique du tout-venant ». La

[2] LABOV, W., & ENCREVE, P. (1976). *Sociolinguistique*.

linguistique de corpus ne nous semble intéressante qu'à partir du moment où elle s'inscrit dans la démarche réflexive de la sociologie qui prône une rupture avec les analyses spontanées :

> *« La sociologie ne peut se constituer comme science réellement coupée du sens commun qu'à condition d'opposer aux prétentions systématiques de la sociologie spontanée la résistance organisée d'une théorie de la connaissance du social dont les principes contredisent point par point les présupposés de la philosophie première du social » [Bourdieu, Chamboredon, Passeron 68 :30[3]].*

Et en premier lieu celles reposant sur un monde social transparent dont il suffirait de prendre en compte les productions attestées comme le critiquait déjà Durkheim lisant Marx :

> *« Nous croyons féconde cette idée que la vie sociale doit s'expliquer, non par la conception que s'en font ceux qui y participent mais par des causes profondes qui échappent à la conscience [Durkheim 1895 :149][4].*

Le projet est bien autre :

> *« Du fait que, à l'occasion de l'observation ou de l'expérimentation, le sociologue entre dans une relation avec son objet qui, en tant que relation sociale, n'est jamais de pure connaissance, les données se présentent à lui comme des configurations vivantes, singulière et, d'un mot, trop humaines, qui tendent à s'imposer comme structures d'objet. En mettant en pièces les totalités concrètes et patentes qui sont données à l'intuition, pour leur substituer l'ensemble des critères abstraits qui les définissent sociologiquement – profession, revenu, niveau d'instruction, etc.-, en interdisant les inductions spontanée qui par un effet de halo, conduisent à étendre à toute une classe les traits marquants des individus les plus « typiques » en apparence, bref en déchirant le réseau de relations qui se tissent continument dans l'expérience, l'analyse statistique contribue à rendre possible la construction de relations nouvelles, capables, par leur caractère insolite, d'imposer la recherche des relations d'un ordre supérieur qui en rendraient raison. [Bourdieu, Chamboredon, Passeron 73:28[5]].*

Toutefois cette posture d'une sociologie résolument en rupture avec le réel, qui prône l'objectivation, ne pouvait prendre en compte le biais possible qui apparaitra

[3] BOURDIEU, P., PASSERON, J.-C., & CHAMBOREDON, J.-C. (1968). *Le métier de sociologue*.
[4] DURKHEIM, É. (1895). *Les règles de la méthode sociologique / par Émile Durkheim,...*
[5] BOURDIEU, P., PASSERON, J.-C., & CHAMBOREDON, J.-C. (1968). *Le métier de sociologue*.

par la suite avec la puissance de la domestication (au sens de Goody) induite par les pratiques du numériques.

Empirie, corpus conducteurs et sorties de route *[retour]*

Le second malentendu peut venir d'une dévotion à l'empirie qui rejoint celle des données dans la perspective « *corpus driven* » et/ou d'une démarche empirique. Là aussi il ne s'agit pas de réduire l'adéquation observationnelle à une empirie dénuée de réflexivité sociologique.

Dans la perspective d'une linguistique « conduite par les données » en opposition à une linguistique basée sur des corpus réservoirs d'exemples, les données préexistent au travail du chercheur et il convient de les analyser en ne mobilisant aucun apriori théorique.

Cette conception nous semble excessivement naïve quant au travail de collecte et de structuration des données qui forment les corpus. Il est fréquent de rencontrer des analyses fondées sur des corpus gigantesques (la Toile) ou une ressource considérée comme représentative par le simple fait qu'elle préexiste à l'investigation. S'il nous semble éminemment pertinent de construire une linguistique à partir des données, cela n'est possible qu'à partir d'une véritable maitrise et réflexion sur la phase de construction de celles-ci.

Là encore la linguistique doit nécessairement beaucoup à la sociologie :

> *Poser avec Bachelard que le* fait scientifique est conquis, construit, constaté*, c'est récuser à la fois l'empirisme qui réduit l'acte scientifique à un constat et le conventionnalisme qui lui oppose seulement le préalable de la construction. [Bourdieu, Chamboredon, Passeron 73:24*[6]*]*

Primauté de l'enquête linguistique VS corpus primaires *[retour]*

On l'aura compris il ne s'agit pas de défendre une linguistique de corpus qui ne poserait pas la question des données et de l'enquête linguistique comme acte fondateur d'une science de la langue. C'est bien parce que la variation relève de la nature sociale de la langue qu'il est nécessaire de disposer de corpus de données sociologiquement construites.

En effet, le matériau linguistique est par essence variable. Or, cette variabilité doit être appréhendée à partir de données situées :

> *« Il suffit de rétablir au premier plan ce qui crève les yeux : que le social s'il est uni est aussi divisé, qu'il est le champ de contradictions*

[6] Ibid.

> *et d'affrontements, et que la langue (comme système, structure, machine) est partie prenante et partie prise de ces divisions ». (Encrevé 1976 :12[7])*

L'affirmation d'Encrevé a une incidence très forte sur une méthodologie de collecte de données linguistiques : celles-ci ne peuvent être dissociées d'une sociologie qui se manifeste à la fois dans la méthodologie de l'enquête et dans la science de la construction de l'observable. Un corpus ne peut se réduire à une simple collection de faits linguistiques. Cette collection est ordonnée, construite sous la responsabilité du chercheur, ce qui nécessite une approche réflexive à l'inverse d'une approche purement empirique ou construite sur des données non théorisées.

Il serait erronée de penser qu'il s'agit ici d'une approche purement sociologique. Comme le relève Depecker :

> « Ainsi la langue n'est pas un être abstrait ou un pur système qui système qui aurait en lui-même sa loi d'évolution et se développerait dans une sorte d'empyrée inaccessible aux mortels. Expression de la « masse parlante », elle est toute entière traversée par les forces sociales. Saussure affirme avec force « *le fait avant tout social de la langue* » (Note de phonologie, 1897, Écrits, p247). Il faut ici prendre en compte toute la portée de cette phrase. *Avant tout* : avant toute autre caractérisation. » (Depecker 2009 :137) [8]

0.2 Principales caractéristiques du parcours de recherche exposé [retour]

Après avoir écarté ces risques de malentendus si fréquents dans un domaine qui n'en est pas un mais qui est en donne l'illusion pour s'être donné une dénomination, celle de *linguistique de corpus*, nous pouvons préciser les caractéristiques qui jalonnent le parcours suivi dans ce travail.

Quel objet scientifique ? [retour]

La première caractéristique est celle qui définit l'objet scientifique. L'ensemble de ce travail repose sur une réflexion sur la construction de corpus oraux. Pourquoi oraux ? Résolument pour ancrer la linguistique dans une discipline son-sens. La linguistique en général, et la linguistique française en particulier, a une tendance forte à travailler sur l'écrit en oubliant qu'il s'agit alors d'une donnée secondaire qui est déjà une annotation formalisée et normalisée.

En effet, si la variation reste toujours présente à l'écrit c'est bien dans la forme orale de la langue que l'on découvre sa nature sociale dans toute son étendue. Or cette nature sociale n'est appréhendable qu'à partir de pratiques linguistiques sociologiquement situées. Il en

[7] LABOV, W., & ENCREVE, P. (1976). *Sociolinguistique*.
[8] DEPECKER, L. (2009). *Comprendre Saussure: d'après les manuscrits*.

découle que l'objet scientifique de la linguistique, celui qui se situe au sein du circuit de la parole selon Ferdinand de Saussure quand le matériau intrinsèquement variable se stabilise en langue partagée, n'est pas dissociable des conditions de sa production et de sa réception au sein de marchés linguistiques qui ne sont observables qu'à partir de véritables enquêtes linguistiques au sein desquelles la sociologie vient à la rencontre de la linguistique.

Les bouleversements technologiques : saisir et/ou produire la variation ? *[retour]*

La seconde caractéristique est la reconnaissance du rôle de la technologie dans l'histoire de la linguistique. Ainsi la linguistique fondée sur l'oral a été bouleversée au début du vingtième siècle quand l'avancée technologique permit l'enregistrement de la parole.

Ferdinand Brunot, dont on retrouvera fréquemment l'inspiration dans cet essai le pressentait dès le début du XX^{e} siècle :

> *« Il suffit qu'une voix s'éteigne pour que nous en soyons séparés par un espace infranchissable, mais nous sommes au siècle des merveilles, si le monde a comme on le dit des réalités que la science n'atteint pas, en échange la science donne sans cesse au monde des réalités qu'il n'avait pas. C'en est fait des lieux communs que depuis l'Antiquité on répétait sur la parole ailée ou sur l'homme attaché à la terre, voici qu'à peu près en même temps l'homme commence à faire son chemin vers le ciel, la parole se grave dans la matière pour toujours ». (Brunot 1911[9])*

ark:/12148/bpt6k1279113

Antoine Meillet [Meillet 1913] l'a souligné, Brunot avait compris que l'archivage non graphique de la parole représentait une ouverture considérable vers la variation car, contrairement à l'écrit, ce que l'enregistreur capte et conserve ce sont les pratiques hétérogènes des locuteurs au sein de leurs productions quotidiennes. Il importerait alors de « démêler le fait accidentel et individuel d'avec le fait général et permanent"10.

Plusieurs décennies plus tard, c'est la révolution numérique qui transformera l'objet scientifique. Abordée tout d'abord par la question de la conservation et de l'archivage, la numérisation de l'analogique ouvrira rapidement un continent à la recherche tant l'objet d'étude s'en trouvait transformé. En premier lieu, les outils qui tamisent et uniformisent

[9] Enregistrement audio : http://gallicalabs.bnf.fr/ark:/12148/bpt6k1279113
[10] Pierre Encrevé 1988, La liaison avec et sans enchaînement, Le Seuil, Paris, p 13, citation de Meillet 1913 et Brunot 1913

l'hétérogénéité linguistique en objet domestiqué, modifieront considérablement l'objet. Plus que jamais, les données sont construites et les acteurs ne sont pas équivalents à de pures technologies. Ce sont également des agents et des faits sociaux qu'on peut facilement cataloguer, voire analyser.

Il faut néanmoins préciser que s'il est encore trop tôt pour évaluer précisément l'impact des sciences du traitement de l'information numérique sur les sciences humaines et sociales, il est grand temps de produire une démarche réflexive sur les corpus, produits de la rencontre entre ces deux domaines scientifiques. Ainsi, si l'enregistrement de la parole bouleverse la méthodologie scientifique en ménageant un nouvel accès à l'objet scientifique dans toute ses variations, les pratiques liées au numérique sont d'un autre ordre du fait de la transformation systématique de l'objet qui se trouve fortement conditionnée par le point de vue.

Les chapitres consacrés à la transcription et aux métadonnées aborderont d'une manière très concrète une nouvelle forme de variation produite par les choix méthodologiques et, en l'occurrence, technologiques du chercheur.

Corpus dédié VS corpus partagé et construction de la connaissance *[retour]*

La troisième caractéristique concerne la prise de conscience de la transformation d'un corpus qui, de corpus d'étude, se transforme en un corpus collecté et construit en prenant en compte un objectif de partage et d'interopérabilité.

Un corpus peut-il exister en dehors d'une étude spécifique ? En 2015 la réponse semble évidente et le débat tranché. Lors de la création de la Très Grande Infrastructure de Recherche en Humanités numériques, Huma-Num, les linguistes se sont organisés en groupes de travail dans deux consortiums avec des objectifs clairement affichés de standardisation des données et des métadonnées afin de définir les conditions de la réutilisation et de l'interopérabilité des corpus. La tendance est générale ; elle se décline au niveau européen dans les infrastructures de recherche CLARIN et DARIAH.

Les linguistes ont toutefois, moins que d'autres disciplines, théorisé les principes qui guident cette orientation.

On peut néanmoins en repérer certains :

- l'aspect économique. Les corpus de langues coutent très chers dès qu'ils sont transcrits et/ou annotés. Il devient alors dispendieux de les produire pour une étude spécifique et ils paraissent hors de portée d'un chercheur ou d'un projet.

- L'outillage des corpus. Les travaux d'Habert [1997-2005] ont démontré et interrogé le rôle d'une linguistique outillée qui ne peut relever du travail d'un seul chercheur et pour une seule finalité.

Par-delà ce simple outillage, c'est la puissance d'une connaissance fondée sur les relations entre les données qui sera interrogée, notamment dans le chapitre sur le web sémantique, mais d'une manière plus générale sur la réflexion qu'appelle la construction de la connaissance scientifique.

- le mouvement des logiciels et des données libres / ouvertes. Même si là aussi la réflexion semble peu maitrisée par les chercheurs entre un modèle qui prône la gestion communautaire des données dans un circuit d'échange (le monde du libre) et un modèle de non propriété des données (*l'open data*), les attendus économiques et sociaux sont saisis dans une dynamique proche d'un certain militantisme.

- L'interdisciplinarité. Les corpus sont appréhendés comme un objet observable depuis de nombreux points de vue qui dépassent ceux du domaine scientifique d'origine.

Rendre les données, objectiver l'institutionnel *[retour]*

La quatrième caractéristique qui a orienté une vingtaine d'années de travail de recherche, concerne la posture du chercheur en sciences humaines et sociales.

La conscience que des données ne sont jamais « données » dépasse le cadre scientifique et concerne la place que prennent les données dans l'espace public. Les corpus oraux, pourtant constitués majoritairement de paroles issues de locuteurs acceptant de confier leurs productions à la science, ont été rarement acceptés comme des objets patrimoniaux. Il convient pourtant de les conserver, les archiver, les exposer et les réutiliser non seulement comme tout objet que la science constitue, mais également comme tout objet reconnu comme un élément du trésor commun. Si cette démarche semble évidente à tout un chacun quand il s'agit de livres, il est loin d'en être de même quand il s'agit de données linguistiques, a fortiori de données orales.

Ces questions croisent deux préoccupations : celle de la valorisation de la démarche scientifique et celle de la politique linguistique.

J'aborderai, dans cette présentation des travaux, à partir d'une approche réflexive, la description d'un certain nombre de réalisations concrètes qui ont pour objectif d'aborder les corpus comme un enjeu de politiques scientifiques mais aussi sociales, culturelles et éducatives.

Cette présentation nécessite d'objectiver la place des contraintes et conduites institutionnelles dans une activité de recherche. La création d'une Agence Nationale de la Recherche, un mouvement très fort d'orientation de la recherche sur le modèle d'un fonctionnement en projet(s), l'apparition d'Idex, Labex, Equipex, de structures d'évaluation de la recherche ont également un impact sur les objectifs d'analyse mais aussi sur le mode de constitution des données de la recherche. Nous verrons jusqu'à quel point cette nécessité d'objectiver prend en compte les effets de variation induits par les institutions. Les conséquences induites par une approche interdisciplinaire telle que celle présentée supra sont un exemple d'une démarche particulièrement encouragée dans le monde de la recherche au début du XXI^{e} siècle et dont les effets sont lisibles au niveau des données (Cf. un exemple pertinent dans [Filhon 2013][11]).

De l'absolue nécessité d'une démarche réflexive *[retour]*

La cinquième caractéristique du parcours de recherche présenté concerne la démarche réflexive qui, du fait du bouleversement de la relation aux objets scientifiques, aux pratiques des chercheurs et à l'accroissement des interfaces avec les autres agents devient une absolue nécessité.

Comme nous l'avons évoqué, la linguistique de corpus présentée ici nécessite une véritable science de la constitution de l'observable. Il s'agit d'une activité onéreuse, chronophage qui engage souvent plusieurs années d'investissement pour des chercheurs et des équipes. A une démarche par étapes cumulatives sous forme d'une chaine de traitements ou de *process* stratifiés, particulièrement risquée, on peut opposer une démarche s'appuyant sur une réflexivité systématique et donc un *process* en *bootstrapping* permanent. Celle-ci a pour avantage de ne pas séparer les opérations de collecte de données et d'analyses, les unes techniques, les autres scientifiques.

Nous considérons les phases de collecte et de traitement des données comme relevant d'une démarche scientifique assumée, y compris quand elle souhaite donner toute sa place à l'empirie. De même, les analyses ne peuvent s'affranchir d'une prise en compte des conditions de production et de traitement des données qui les forgent et dont les cadres théoriques sont parfois fortement implicites.

Il n'y a donc pas une chaine de traitement en linguistique de corpus, chaine dont l'analyse serait le point d'orgue ultime. Le linguiste, en tant que scientifique, se doit d'être présent du début à la fin dans cette démarche dont les aspects techniques, technologiques, méthodologiques et théoriques sont imbriqués dans une démarche scientifique unique. Plus

[11] FILHON, A. et al. (2013). « Un projet de nomenclature socioprofessionnelle européenne ».

que jamais il se doit d'avancer pas à pas en répondant systématiquement à l'injonction : « que le linguiste sache ce qu'il fait ».

A la croisée du scientifique et du politique [politique linguistique] *[retour]*

La sixième caractéristique de mon travail est la reconnaissance du parcours particulier qui consiste à mener de front un travail de constitution et d'analyses d'un corpus et des activités de politique de la recherche au sein du Ministère de la Culture et de la Communication (Délégation Générale à la Langue française et aux Langues de France) et au MESR. Ces activités ne sont pas séparées ou parallèles, elles forment un tout scientifique à partir du moment où l'on adopte une relation obligatoire et nécessaire entre une approche théorique qui ne renie pas la nature sociale de son objet et les enjeux d'une utilisation de ces données et analyses à des fins de politiques sociales, culturelles, éducatives tout autant que scientifiques.

Le corpus comme production scientifique *[retour]*

La dernière caractéristique de ce travail, *in fine* la plus essentielle et aussi la plus discutée, est de plaider lourdement pour la reconnaissance du travail de constitution de corpus en tant que production scientifique à part entière. Ainsi l'objectif de cette présentation est d'exprimer une synthèse des travaux réalisés depuis la thèse dont une part considérable a été investie dans la collecte et la diffusion d'un très grand corpus oral (ESLOs, composé d'une enquête réalisée en 1968-74 : ESLO1 et d'une enquête en cours de réalisation depuis 2004 : ESLO2) et dans la conduite d'un vaste programme de compilation de corpus dans plusieurs dizaines de langues représentant plusieurs milliers de documents : le programme *Corpus de la parole*.

Trois arguments appuient ce plaidoyer. Premièrement, le travail de collecte et de gestion du corpus est systématiquement relié à une démarche théorique et une approche réflexive des phases méthodologiques. En ce sens, l'activité même de collecte est totalement conçue comme une activité scientifique. Deuxièmement une part très importante de l'énergie consacrée à la gestion du corpus a été investie dans sa diffusion selon des « bonnes pratiques » afin d'en faire un objet disponible pour d'autres études scientifiques, en partage dans une vaste communauté. Troisièmement les deux corpus qui seront présentés forment une masse considérable de données et donc un travail en « back office » des plus fastidieux. Le Corpus des ESLOs, dans sa forme immédiatement disponible sur les plateformes de diffusion au 14 juillet 2015 représente plus de 500 heures d'enregistrements soit plus de 7 millions de mots transcrits. Il est composé de 599 enregistrements, 1428 fichiers de transcription, 2027 fichiers de métadonnées et 628 documents complémentaires (fiches descriptives, autorisations, carnets de terrain). Il s'agit de la forme diffusée du corpus mais

une brève introspection de mes espaces de travail donne un aperçu concernant l'ampleur du travail de préparation : l'espace de stockage de « production » contient 9827 fichiers et le dossier ESLO de mon espace de travail personnel, qui ne contient ni enregistrements, ni transcriptions mais seulement des documents de travail (comptes rendus de réunion, demande de financement, suivi de projets, tableaux d'analyse etc.) comprend 7558 fichiers dont le plus ancien date du 7 avril 2004. Bien évidemment il ne s'agit pas de m'attribuer la totalité du travail sur ce corpus et ma part personnelle varie considérablement. Ainsi, dans ESLO1 j'ai un rôle de «compiler» sur l'ensemble du corpus et de chercheur sur l'ensemble des transcriptions sans que j'aie réalisé d'entretien ou de transcription. Pour ESLO2, il en est autrement puisque j'ai à la fois un rôle de responsable du projet, ce qui me fait intervenir intensément à tous les niveaux, et j'ai collecté personnellement une part des enregistrements. Un dernier chiffre peut permettre d'appréhender une réalité difficile à objectiver. Depuis 2004, j'estime mon temps de travail de collecte et de gestion du corpus à une moyenne 12h/semaine soit plus de 6 000 heures de travail. Si, à titre de comparaison sommaire, on estime qu'un article de recherche de qualité moyenne représente environ 200 heures de travail, le temps passé sur le corpus ESLO est comparable à plus de 30 articles.

L'estimation du temps passé à construire le programme Corpus de la parole doit être appréhendé autrement et mon rôle aura été différent mais je peux estimer grossièrement le temps passé à environ la moitié de celui consacré à ESLO, ce qui représenterait 3 000 heures de travail soit une quinzaine d'articles.

Plus simplement on peut estimer que la constitution et la gestion du corpus représentent 9 000 heures de travail consacrées à ces productions scientifiques.

0.3 Plan du document *[retour]*

Plan

0. Introduction
1. De l'expérience personnelle à un parcours de recherche
[Ce chapitre a pour objectif de relever ce qui dans le parcours antérieur à la recherche ou dans les activités parallèles mais externes éclaire les chemins empruntés dans le cadre des travaux de recherche]

2. Légitimité de l'enquête sociolinguistique et des corpus)

[Ce chapitre vise à décrire brièvement quelques cadres théoriques. Il s'agit notamment d'en repérer les passerelles et les contradictions]

3. ESLO

[Ce chapitre est résolument descriptif. Il présente l'activité de constitution d'un corpus sur l'ensemble de la chaine, de la collecte à l'analyse, afin d'éclairer ce qu'une méthodologie détaillée et explicite peut offrir comme pistes d'analyse. La dernière partie du chapitre souhaite démontrer sur un exemple d'analyse les relations méthodologies/théories]

4. Politique linguistique et politiques de recherche

[Ce chapitre, lui aussi résolument descriptif, présente l'ensemble de mes activités réalisées dans le cadre de la DGLFLF. Il s'agit de montrer le lien fort qui existe entre politique linguistique et science du corpus et des effets que l'un a sur l'autre et réciproquement.]

5. Conclusion

0.4 Présentation du document d'HDR *[retour]*

Cette synthèse de travaux, rédigée dans le cadre d'une Habilitation à Diriger des Recherches, présente trois spécificités de forme qui sont des éléments assumés de la narration d'une activité scientifique. Il s'agit également d'une tentative de réflexivité poussée jusque dans les applications au rendu même de l'observation sous forme de production scientifique des caractéristiques de l'objet observé.

1. Contrat de lecture *[retour]*

Cette HDR se confronte directement à la notion de « texte linéaire » et à la pertinence de l'usage de celui-ci pour appréhender une réflexion scientifique fondée sur différentes activités, projets, intérêts qui sont parfois distincts, parfois redondants et dans des temporalités qui peuvent être autant synchrones que disjointes. Le choix du rédacteur s'est porté vers la notion de « contrats de lecture » afin de mettre en évidence différents parcours de lecture qui s'appuient sur une gestion de l'information dont l'architecture construit un sens différent de celui que produirait une lecture linéaire de la somme de ses parties.

Ainsi nous utiliserons des cartes heuristiques pour permettre différents parcours de lecture tout en essayant de mettre en avant la cohérence générale de la réflexion. A une carte générale s'ajoutent des cartes par chapitres ou grand thèmes.

Afin de tirer profit au mieux de la flexibilité des parcours de lecture, le texte est découpé en blocs indexés qui sont reliés par des hyperliens dans toute l'HDR. Liens internes
Enfin des liens vers des URLs permettent de donner accès directement à chaque fois que c'est possible à du contenu présent sur la Toile. HDR sur Nakala

Cette tentative de « contrat de lecture » est également une façon de s'inscrire dans une réflexion sur la gestion de l'information « éclatée », pour laquelle il appartient de plus en plus au lecteur de développer une compétence nouvelle afin de construire le parcours le plus pertinent comme l'évoquent Salaün et Habert :

> *Un fil, une thématique implicite, une clé traverse donc l'ensemble du livre et relie ses chapitres : la conviction d'assister à une transformation de l'ancien contrat de lecture en un contrat de lecture/écriture renouvelé, transformation où les architectes de l'information occupent une place centrale. Sous l'expression* contrat de lecture*, on réunit un ensemble de conventions implicites ou explicites de forme, de contenu ou encore de transmission, qui assure la communication entre l'entité émettrice (auteur, producteur, éditeur, etc.) et l'entité réceptrice (destinataire, lecteur, spectateur, utilisateur, client, etc.). Grâce à ces conventions, le premier (émetteur) s'adresse effectivement au second (destinataire) et ce dernier est capable 1) de repérer, de classer dans un genre un document, 2) de le lire ou le décrypter, lui donner du sens, 3) de s'approprier le contenu, c'est-à-dire de modifier son comportement à partir des éléments proposés par le premier (émetteur).* [Salaün et Habert, 2015:186][12]

Le lecteur pourra donc se laisser guider soit par une lecture linéaire en lisant du début à la fin le texte dans l'ordre des chapitres, soit en suivant une lecture thématique qui relie différentes informations produites dans différentes expériences de production scientifique. Ainsi une analyse du rôle et de l'effet du recueil de consentement dans la méthodologie de terrain pourra être abordée à partir de l'expérience d'un groupe de travail national sur les bonnes pratiques juridiques mais aussi à partir de l'expérience directe d'un terrain d'enquête ou encore à partir des enjeux de la structuration d'une recherche transdisciplinaire.

[12] SALAÜN, J.-M., HABERT, B., & COLLECTIF. (2015). *Architecture de l'information : Méthodes, outils, enjeux*.

2. Accès direct à la production scientifique « disponible ». *[retour]*

Il ne s'agit pas seulement de donner accès aux documents en cohérence avec un « mouvement » pour les données ouvertes mais d'apporter un regard réflexif et parfois prospectif sur trois points :

- permettre un accès aux données, en l'occurrence les productions scientifiques du chercheur, afin que cette synthèse des travaux dépasse une méta-présentation et applique sur tout objet une méthodologie d'inclusion des données situées dans l'analyse, qui constitue le cœur de l'activité scientifique,

- tenter de définir un parcours de recherche à partir d'une cartographie sur la Toile des productions scientifiques d'un chercheur, lui-même se présentant comme agent situé,

- observer les usages possibles d'une production scientifique qui échappe au chercheur et qui se trouvera réagencée par le parcours du lecteur en fonction également de l'enrichissement fourni par les outils d'exploitation des données sur la Toile.

- Typologie des contenus relevant d'une production scientifique accessible sur la Toile :
  - Publications

    Celles-ci sont accessibles par la plateforme d'archives ouvertes du CNRS HAL. Ces productions disposent d'un identifiant unique et de la référence de leur publication. Dans ce texte, j'utiliserai la référence bibliographique pour les citations ou renvois et l'URL quand il s'agira de donner directement accès à la ressource.

  - Corpus
    - Enregistrements et transcriptions

    Les enregistrements sonores et les transcriptions sont archivés à l'aide du service géré par la plateforme COCOON. Ils sont également disponibles sur la plateforme de L'EQUIPEX ORTOLANG et par d'autres moyens qui seront présentés et discutés dans le cadre de ce travail.

    Conformément aux pratiques en vigueur, chaque document dispose de métadonnées et d'un identifiant unique. J'utiliserai cet identifiant pour citer la ressource et l'url quand il s'agira de donner directement accès aux données.

    - Documents liés à la méthodologie de corpus

La production de corpus nécessite également des productions scientifiques qui accompagnent l'ensemble de la démarche : protocole, guide de transcription, texte de vulgarisation. Quand ceux-ci n'ont pas donné lieu à une publication, ils ont été déposés et exposés à l'aide du service Nakala de la TGIR Huma-Num. Ils disposent ainsi d'un identifiant unique et d'une URL y donnant accès.

- Observatoire des pratiques linguistiques et politique linguistique

Outre des publications classiques, souvent dans des revues non scientifiques, mon activité au sein de la DGLFLF donne lieu à la production de différents documents : texte de présentation d'appels à projets, compte rendu de conseils scientifiques etc.
Quand ces documents ne sont pas disponibles parce le Ministère de la Culture n'a pas encore mis en place une réelle politique d'archivage appliquée aux documents de son administration, ceux-ci ont été déposés dans un espace Nakala dédié aux travaux de cette HDR.

- Documents divers

  Enfin certains documents ne relevant d'aucune des autres catégories ont également été déposés dans l'espace Nakala afin de disposer d'une procédure d'archivage pérenne et d'un identifiant unique qui, liés à une url, permettra à chacun d'avoir un accès à ces données selon ce processus de publication.

3. Présentation des corpus *[retour]*

Comment décrire et présenter un corpus comme objet d'une production scientifique ? Cette question n'admet pas de réponse évidente et elle se pose de façon cruciale pour ce travail.
Le premier réflexe, conforme à cette approche, serait d'utiliser un format de métadonnées, en l'occurrence celui utilisé en premier lieu pour les corpus évoqués ici, le DUBLIN-CORE qualifié.
Les 15 descripteurs sont les suivants: (Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type)[13].
Nous le verrons, cette phase de description des corpus est encore balbutiante et ne répond pas à tous les objectifs, et en premier lieu pas à celui qui nous intéresse dans le cadre d'une présentation synthétique d'un parcours de production scientifique.

Titre
La première information relève d'une action de dénomination.

[13] http://dublincore.org/

Il est important de décrire un corpus comme un objet relevant d'une unité structurée. En ce sens, un corpus doit avoir un titre, même et surtout s'il est composé d'une collection de documents qui ont eux-mêmes leur propre identité.

Auteur

La seconde information importante est celle de l'auteur de la ressource. Il s'agit d'une question complexe que les aspects juridiques aident à définir mais qui dépasse largement ce cadre pour couvrir celui de l'éthique, de l'économique et du sociétal. Ce point est d'autant plus complexe qu'il y a souvent de nombreux auteurs intervenant à différents niveaux dans la vie d'un corpus. Dans cette HDR où la notion d'évaluation qualitative mais aussi quantitative de la production scientifique est présente, je proposerai une estimation subjective à l'aide d'un pourcentage de la part d'auctorialité que je pense être en mesure de revendiquer à chaque fois que ce sera possible et/ou nécessaire.

Type

La troisième information concerne le type de données présentes dans le corpus. Il convient de préciser s'il s'agit de données orales/multimodales ou de transcriptions ou d'autres formes d'annotation secondaires.

Structure

La quatrième information est celle de la structure du corpus. Nous avons retenu la collection de documents comme définition du corpus, or la qualité de celui-ci dépend fortement de la richesse de la structure de cette collection qui ne peut se réduire à un amoncèlement quantitatif et linéaire de données.

Taille

La cinquième information concerne l'évaluation de la taille du corpus. Là non plus il n'est pas simple de fournir une information pertinente. Dans le cas des corpus linguistiques, il est d'usage d'évoquer le nombre de mots comme échelle quantitative. Celle-ci est pourtant insuffisante et totalement différente selon le type de données primaires (orales/multimodales ou « nativement » sous forme d'écrits numériques). Nous utiliserons plusieurs critères dans le cadre d'un corpus oral : le nombre de documents, la durée de l'enregistrement, le nombre de mots de la transcription et le cas échéant la taille du fichier numérique.

Projet

La sixième information concerne la description des objectifs du corpus et sa relation avec un projet scientifique situé. Cette information est primordiale et pourtant force est de constater qu'elle est très souvent implicite voir non disponible.

Date et version

La septième information doit permettre de situer le corpus dans sa période de collecte et de production. Elle permet aussi de préciser les versions et les relations entre celles-ci afin de tracer l'évolution du corpus.

Licence et accès

Enfin, la huitième information concerne les conditions d'accès et de diffusion du corpus. Cette information est déterminante dans une optique de contrôle de la falsification des données. S'il est tout à fait compréhensible qu'un corpus, pour des raisons juridiques ou éthiques, ne puisse être accessible au grand public, il est par contre délicat de concevoir qu'il ne soit pas accessible dans un cadre purement scientifique.

1. Curriculum Vitae et parcours personnel [retour]

Parcours personnel

Il aurait été possible d'appliquer à l'auteur de ce document, la méthode de description sociologique utilisée dans le cadre du programme des ESLO en situant un individu au sein des réseaux, pratiques et activités d'une ville. Je me contenterai de donner quelques éléments contextuels…

Je suis né à Orléans, septième enfant issus d'un père cadre aux Impôts natif de Boulogne-sur-mer et d'une mère enseignante de français née à Moulins. Le modèle familial est assez fréquent. Les parents mariés à la fin de la guerre, catholiques de gauche, s'orientent après 1968 vers un dynamisme militant à la fois dans le milieu politique, syndical, associatif et culturel.

C'est ainsi que j'ai grandi en participant dès mon enfance aux évènements militants locaux avec les mêmes participants que dans les activités culturelles : l'association du ciné-club, celle du théâtre, « la » librairie, les évènements politiques… Le moule est assez classique et au niveau de granularité d'une ville de province moyenne, les réseaux de connaissances et d'activités sont particulièrement lisibles.

Tout au long de mes études universitaires j'ai gagné ma vie en tant qu'animateur puis directeur de centres de loisirs, principalement dans les structures de la ville d'Orléans. J'y ai énormément appris, sur la ville et les êtres qui y vivent et j'y ai développé des compétences pour le travail collectif. [mini histoire sociale]

Etudes secondaires et universitaires

Après un parcours dans les écoles et établissements secondaires publics de la ville d'Orléans j'ai obtenu un bac littéraire avec option scientifique. J'ai ensuite suivi une formation en lettres modernes puis en linguistique.

1992 : Maitrise de Lettre Modernes, Université d'Orléans.

1993 : DEA de Linguistique Université Paris 8

C'est au cours de la maitrise que mon gout pour les sciences du langage et la recherche en linguistique s'est développé très fortement. J'ai alors rédigé un mémoire de maitrise à partir

d'une enquête sociolinguistique, *Le sens en scène*, puis un mémoire de DEA en sociopragmatique à partir d'un corpus de presse écrite (mention très bien). Ces deux travaux doivent beaucoup à la lecture de l'article de Pierre Encrevé et Michel de Fornel « le sens en pratique »[14].

J'ai été très fortement marqué par la pensée de Pierre Encrevé ainsi que par les qualités humaines de celui-ci. Cette influence sur mon travail a été et reste considérable.

Les séminaires suivis à l'EHESS pendant une dizaine d'année m'ont également procurés une joie immense et très formatrice.

Thèse

C'est ainsi qu'en 1998 j'ai soutenu une thèse en sociopragmatique cognitive à l'EHESS sous la direction de Pierre Encrevé : *Le sens sous presse. Une approche cognitive et sociologique de la construction du sens d'un terme lexical au cœur d'un évènement médiatique, un exemple: la réforme de l'orthographe de 1990*.

Résumé court :

La polémique dans la presse écrite qui a entouré "*la réforme de l'orthographe de 1990*" fut l'occasion de construire à partir d'un corpus de presse, un objet permettant de mener à bien des recherches en sociopragmatique cognitive. En effet, une analyse a posteriori de l'évènement laissait l'impression qu'il avait donné lieu à une vraie lutte politique, cognitive mais surtout sémantique: une lutte qui a abouti à une modification de la représentation d'un acte par la modification du sens des termes utilisés pour désigner celui-ci. Un rapport technique recommandant des "rectifications utiles et modérées" s'est transformé au cours d'une vive polémique en une "réforme profonde de l'orthographe" accusée de mettre la langue française en péril. Comment le sens du terme désignant l'évènement a-t-il évolué? Comment a-t-il dérivé ? Comment s'est-il construit à la fois et simultanément au sein d'une communauté linguistique et au cœur de chacun des membres de cette communauté ? Pour répondre à ces questions il convenait de conjoindre des concepts opératoires issus de la sociologie de Pierre Bourdieu, de la sociolinguistique de William Labov et de la sémantique et de la pragmatique des « espaces mentaux » et de « l'intégration conceptuelle » de Gilles Fauconnier et Mark Turner.

Cette thèse a été validée avec les félicitations du jury à l'unanimité. La suite de mon parcours de recherche est décrite et questionnée dans cette présente synthèse.

[14] ENCREVE, P. et al. (1983). *Actes de la recherche en sciences sociales. L'usage de la parole*.

ENSEIGNEMENT :

En 2000 j'ai été recruté comme Maitre de conférences en linguistique à l'Université d'Orléans. Après trois ans d'enseignement en IUT au département *Techniques de commercialisation* j'ai rejoint le département des sciences du langage.

Licence :

Introduction à la linguistique
Lecture de textes en sciences humaines
Sociolinguistique
Terrain, Enquête, Corpus,
Psycholinguistique et cognition
Pragmatique de l'interaction
Langage et communication
Linguistique cognitive
Théories de la communication

Master, doctorat :

Techniques d'enquête
Linguistique cognitive
Introduction à la linguistique
Analyse de conversations
Analyse du discours politique et médiatique
Politique linguistique
Psychosociologie de la communication
Prise de parole publique
Rédaction professionnelle

Nombre d'heures en EQTD sur les cinq dernières années :

2011 : 260
2012 : 268
2013 : 277
2014 : 282
2015 : 284

Par vocation et par intérêt intellectuel j'ai toujours souhaité être enseignant-chercheur. Ce statut, obtenu en 2000, m'a très rapidement permis de m'inscrire pleinement dans les différentes missions d'un universitaire. Ce point me parait particulièrement important et il explique que j'ai consacré beaucoup d'énergie à des responsabilités dites « administratives »

dans les premières années de ma carrière. J'ai donc été engagé dans plusieurs responsabilités administratives pendant huit années à la tête de la direction d'un département qui comptait 350 étudiants.

RESPONSABILITES ADMINISTRATIVES

2004-2007 Directeur des études Licence SDL

2007-2011 Directeur du département des Sciences du langage de l'université d'Orléans.

Responsable de la maquette de la licence Sciences du langage, (évaluation AERES A+)

Responsable de l'option Langue des signes

2007-2012 Membre du conseil de l'UFR LLSH

20011-2015 Membre du Conseil scientifique de l'Université d'Orléans

Parallèlement à mes activités d'enseignement j'ai développé un parcours de recherche assez atypique avec une orientation forte vers des domaines qui croisent le Ministère de l'enseignement supérieur et de la recherche et celui de la Culture et Communication :

- Membre du Laboratoire Ligérien de Linguistique UMR 7270
- Depuis 2012 Membre suppléant au CNU
- 2015- Directeur de la TGIR Huma-Num UMS 3598
  2012- 2015 Président du conseil scientifique de la Très Grande Infrastructure de Recherche Huma-Num http://www.huma-num.fr/le-conseil-scientifique

- Depuis 2000- Directeur scientifique de l'Observatoire des pratiques linguistiques (DGLFLF - Délégation Générale à la Langue Française et aux Langues de France, Ministère de la culture et de la communication).
  http://www.dglflf.culture.gouv.fr/observatoire/observatoire_accueil.htm

- Depuis 2013 Expert pour le JPI On Culture Heritage http://www.jpi-culturalheritage.eu/
- Depuis 2012 membre du comité technique EQUIPEX ORTOLANG
  http://www.ortolang.fr/

- Depuis 2015 Membre du conseil scientifique de la MSH Val de Loire

Lors de ce parcours j'ai été amené à endosser la responsabilité de projet et programmes scientifiques :

J'ai été responsable au sein du LLL du projet Enquêtes Sociolinguistique à Orléans (conservation et mise à disposition d'ESLO1 et élaboration d'ESLO2- actuellement 400 heures de français parlé, 700 heures à terme) http://eslo.huma-num.fr/ et depuis 2004 du programme Corpus de la parole DGLFLF-MCC.http://corpusdelaparole.huma-num.fr/ mais j'ai été également :

- Responsable du projet APR-IA ODIL, 2015-2017
- Responsable scientifique du projet « Sémantisation du corpus de la parole », financement Ministère de la Culture et de la Communication, 2015
- Responsable du projet ESLO 3.0 financement MSH-VDL 2014-2016
- Responsable scientifique pour la DGLFLF du projet « Numérisation des corpus en langues de France » sélectionné, financement MCC, 2005-2008
- Co-responsable du projet ANR VARILING, Corpus en SHS 2006.

Ces différentes responsabilités se sont aussi traduites au sein de comités de pilotage :

- Depuis 2006- Membre du comité de pilotage du plan de numérisation du Ministère de la Culture
- 2011-12 Membre du Comité d'appui interministériel pour le Pôle d'excellence dans le domaine linguistique et des traditions orales en Guyane
- Depuis 2013 Membre du comité de pilotage de la Recherche au Ministère de la Culture
- 2007-2010 Membre nommé au Cosla (Comité pour la Simplification du Langage Administratif).
- Depuis 2010 Participant au Réseau européen de la recherche DC-net numérisation du patrimoine Culturel (7e PCERD)

Mes activités de recherche se sont également concentrées sur l'encadrement de thèses et je suis depuis 2015 titulaire de la Prime d'encadrement doctoral et de recherche. J'ai codirigé trois thèses au LLL :

- ANNIE VASLIN-CHESNAULT : ANALYSE DIACHRONIQUE DE LA VARIATION SOCIOLINGUISTIQUE A PARTIR DE DEUX CORPUS ORLEANAIS. 664 pages,

A partir de la comparaison entre deux corpus collectés auprès des mêmes locuteurs à quarante années de distance, il a été procédé à une étude linguistique variationniste. Dans le cadre d'un module de la nouvelle enquête ESLO2 du Laboratoire Ligérien de Linguistique (2002-2010) un échantillon de dix personnes a été constitué en sollicitant des témoins ayant participé à l'Enquête Socio-Linguistique à Orléans (ESLO1 - 1968-1970). Il s'agissait de mettre à l'épreuve l'hypothèse selon laquelle un locuteur conserverait à l'identique sa pratique linguistique au cours de sa vie. Après collecte et transcription (sous Transcriber) des enregistrements, la recherche s'est centrée sur les points suivants : (i)la variation du lexique (catégorisation sociale, technologie, français dit familier) (ii)l'usage des deux éléments de la négation verbale et (iii)les différents types de liaison (enchaînée, non enchaînée, après une pause). Avec le souci de mettre à disposition de la recherche linguistique un corpus constitué suivant les règles de l'art, cette thèse met en évidence la présence d'une variation diachronique chez tous les locuteurs, avec des différences selon l'origine et la trajectoire

Direction : Gabriel Bergounioux, PR (50 %) Co-encadrement : Olivier Baude, MCF (50 %) soutenue en 2009

- LINDA HRIBA M'CHAREK : IDENTIFICATION AUTOMATIQUE DES LOCUS DE VARIATION DANS UN CORPUS DE FRANÇAIS PARLE. Co-encadrement :

A partir d'un travail prospectif sur les locus de variations au sein d'un grand corpus oral, la thèse s'est orientée vers l'analyse des variations captées lors de l'activité de transcription, considérée comme une production de la langue des locuteurs-auditeurs.

Après une présentation de l'état de l'art sur la méthodologie des corpus oraux, sur la place des corpus en sociolinguistique et sur les recherches sur le français parlé, la thèse développe une analyse des variations de transcription repérées dans un sous corpus constitué des trois versions de transcriptions utilisées dans le projet des ESLO.

Direction : Gabriel Bergounioux, PR (50 %) Co-encadrement : Olivier Baude, MCF (50 %)

- ATHENA DUPONT : ANALYSE DES FORMATS ENONCIATIFS EN INTERACTION DIALOGALE

L'étude des documents sonores se fait le plus souvent à partir de l'enregistrement d'interactions entre deux ou plusieurs locuteurs qui sont conduits à construire, dans le dialogue, des types de représentation et de catégorisation de l'interlocuteur et d'eux-mêmes. Cette relation s'actualise de façon évolutive par le jeu de marques formelles qui sont autant d'indices de l'énonciation, que ceux-ci relèvent de paradigmes fermés (pronoms, deixis, tournures interrogatives, TAM…) ou qu'ils émergent comme système dans des usages discursifs inférés par la grammaticalisation et les modalités de référencement des agents. Le recours à des données variationnistes contrôlées, telles que celles recueillies par l'Enquête Sociolinguistique à Orléans (ESLO), doit permettre d'établir un relevé des occurrences et une cartographie des emplois. Sur une batterie d'indicateurs mis à l'épreuve sur corpus, on se propose de mesurer les procédures d'implication et de reconnaissance des locuteurs-auditeurs dans leurs énoncés. Ce travail s'inscrit dans la continuité d'une exploitation systématique des matériaux recueillis par le LLL.

Direction : Gabriel Bergounioux, PR (50 %) Co-encadrement : Olivier Baude, MCF (50 %) Date de début mai 2014.

J'ai également participé à des jurys de thèse :

- Aguilar (2008), Valin-Chesneau (2009), Abourahim (2011), Javier (2011), Kawakami (2012), Lureau (2014), Evora (2015).

Et j'ai encadré des activités post-doctorales :

- 2010 2011 Encadrement des activités de Mme Vaslin-Chesnault dans le cadre du programme de recherche du Laboratoire Ligérien de Linguistique. (Thèse soutenue en 2009, O. Baude Co-encadrant à 50 %).
- 2013- 2015 Encadrement des activités de Mme Layal-Kannan (thèse soutenue en 2011) dans le cadre de l'EQUIPEX ORTOLANG. (CDD de 2 x 10 mois sous la responsabilité d'O. Baude : exploitation scientifique du corpus ESLO, gestion des données).
-

Mes activités d'administration de la recherche et de dialogue scientifique se sont aussi concrétisées dans la participation à des revues :

Comité de rédaction

- 2002-2013 Rédacteur en chef de *Langues et Cité* Ministère de la Culture et de la Communication.
    - 23 numéros

  http://www.dglflf.culture.gouv.fr/publications/publications.htm

- Depuis 2002- Membre du comité de rédaction de la Revue de Sémantique et Pragmatique.
  http://www.univ-orleans.fr/RSP/

- Depuis 2013 Membre du comité de rédaction *Culture&Recherche*, MCC
- Depuis 2010- Membre du Comité lecture des *Nouveaux cahiers de l'AFLS*. http://afls.net/cahiers-e-journal/

Mais aussi dans des expertises :

- Depuis 2013 Expert pour le JPI On Culture Heritage http://www.jpi-culturalheritage.eu/
- 2010 : Deux rapports d'expertise pour le Social Sciences and Humanities Research Council of Canada
- 2014 : Expertise pour l'AAP Conseil Régional Aquitaine.
- Depuis 2010 Participant au Réseau européen de la recherche DC-net numérisation du patrimoine Culturel (7e PCERD)
- 2010 Responsable des expertises pour l'appel à propositions de la DGLFLF Ministère de la Culture : *Alternance codique*
    - 7 dossiers
- 2012 Responsable des expertises pour l'appel à propositions de la DGLFLF Ministère de la Culture : *Numérique et textualité : observation, description et analyse des pratiques contemporaines*
    - 11 dossiers
- 2013 Responsable des expertises pour l'appel à propositions de la DGLFLF Ministère de la Culture : *Observation des pratiques linguistiques en langues de France*
    - 27 dossiers

Enfin, en cohérence avec les pratiques scientifiques en vigueur, je participe à des comités scientifiques et des comités d'organisation de colloques :

Comités scientifiques :

- 2016 Congrès Mondial de Linguistique Française, Tours, session *Ressources et outils linguistiques pour l'analyse de la langue*
- 2015 TALN Atelier Ethique et TAL
- 2015 Colloque SHESL-HTL 2015 - *Corpus et constitution des savoirs linguistiques*

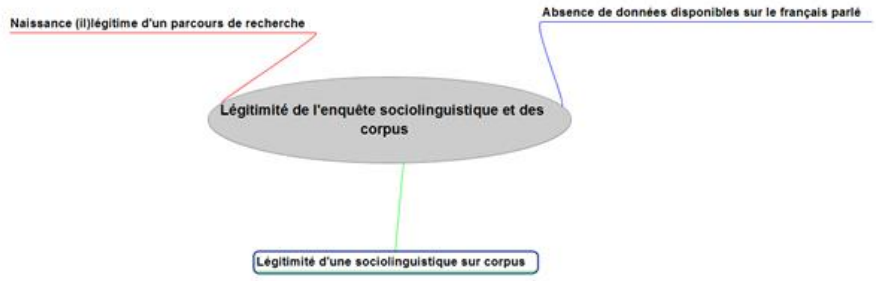
- 2015 Colloque jeune chercheur LLL
- 2015 Colloque RFS, Hétérogénéité et changements : perspectives sociolinguistiques, Grenoble, 10-12 juin
- 2015 Colloque ICODOC, ICAR, Lyon
- 2014 Journée d'études ATALA Ethique et TAL
- 2014 Congrès Mondial de Linguistique Française, Berlin, Session *Ressources et outils linguistiques pour l'analyse de la langue*
- 2014 Colloque international *Les Métropoles Francophones en Temps de Globalisation*, Nanterre, 5-7 juin
- 2011 : Colloque "Corpus, Données, Modèles: approches qualitatives et quantitatives", Montpellier
- 2010 Colloque *Lexique, Normalisation, Transgression,* Cergy
- 2006 RSP4, Quatrièmes Rencontres de Sémantique et Pragmatique, 13 - 15 juin
- 2006 Colloque international de linguistique "Faire signe _ pour Pierre Encrevé" 16, 17, 18 octobre, Paris

Participation à des comités d'organisation de colloques :

- 2015 8es Journées Internationales de Linguistique de Corpus, 2-4 sept Orléans (France)
- 2014 Colloque Jean Zay, 25 et 26 novembre, Paris et Orléans
- 2009 Cinquièmes Rencontres de Sémantique et Pragmatique (RSP5), 22, 23 et 24 avril, Université de Gabès (Tunisie)
- 2006 RSP4, Quatrièmes Rencontres de Sémantique et Pragmatique, 13 - 15 juin
- 2006 Colloque international de linguistique "Faire signe _ pour Pierre Encrevé" 16, 17, 18 octobre, Paris
- 1999 Seconde Rencontres RSP, juin, Orléans

Depuis le premier septembre 2015 je suis directeur de la Très grande infrastructure de Recherche en Humanités Numériques Huma-Num UMS 3598. J'ai néanmoins conservé mon poste à l'université d'Orléans.

2. Légitimité de l'enquête sociolinguistique et des corpus [retour]



2.1 Naissance (il)légitime d'un parcours de recherche [retour]

Le projet des Enquêtes Sociolinguistique à Orléans est indéniablement celui qui a dirigé de manière centrale mon travail de recherche depuis plus de dix ans. Il a un effet structurant sur mon parcours car c'est lui qui a accompagné l'ensemble de ma réflexion sur la sociolinguistique bien évidemment, mais aussi, de manière beaucoup moins prévisible à l'époque du démarrage du projet, vers le tournant des Digital Humanities. On pourra donc lire ce parcours à partir de deux entrées : l'une plus classique consacrée à des analyses sociolinguistiques qui se fondent sur un grand corpus de français parlé et l'autre plus originale qui provient de l'accompagnement méthodologique et technologique de la gestion d'un corpus numérique. C'est bien la réunion de ces deux entrées qui conditionne l'approche réflexive proposée ici et qui oriente l'analyse vers une science de l'observation des données linguistiques dont les conditions de production soient explicites.

Contribution anecdotique à une mini histoire sociale [parcours personnel] [retour]

Mon rapport aux ESLO a été tout d'abord très anecdotique. Après une maitrise de Lettres modernes pour laquelle j'avais rédigé un mémoire consacré à une étude sociolinguistique, je m'étais engagé, avec beaucoup de passion pour la linguistique que je découvrais véritablement à partir des enquêtes sur les pratiques, dans un DEA de linguistique. Lors d'un premier rendez-vous avec l'unique professeur de linguistique de l'Université d'Orléans[15], celui-ci me montra une armoire remplie de 350 cassettes audio rangées dans leur boitier transparent et ornées de jaquettes blanche et rouge, parfaitement alignées sur plusieurs étages et comportant un code mystérieux sur leur tranche. Il s'agissait d'une copie toute

[15] Le jeune professeur Gabriel Bergounioux

fraiche, sur minicassettes, du *Corpus d'Orléans,* enregistré sur bandes magnétiques dans les années soixante-dix.

Je suis ressorti après une heure de discussion avec une fascination à la fois pour la conversation que nous avions eue et qui m'ouvrait la porte d'un continent scientifique et pour cet objet bien éloigné des photocopies d'articles et de livres que j'utilisais depuis le début de ma formation universitaire : un fonds sonore.

Cette découverte réveillait une autre passion dans laquelle je m'étais plongé durant mon adolescence : les enregistrements sonores et la radio. En effet, le grand mouvement des radios libres en France a pris un essor considérable avec la libéralisation des ondes en 1981 et 1982 par le tout nouveau Président de la république François Mitterrand alors que j'avais 14 ans. J'ai ainsi passé beaucoup de temps dans une radio associative de l'agglomération orléanaise à enchainer disques et tribunes dans une vieille maison réquisitionnée, aux murs recouverts de boites à œufs pour parfaire l'isolation phonique. Dans cette radio et dans les quelques autres où je trainais dès que possible, je découvrais des étagères entières de disques et de cassettes qui représentèrent, pour moi, une alternative aux bibliothèques.

Je reprenais d'ailleurs personnellement cette lutte entre deux formes matérielle du savoir en expulsant les vieux livres à tranche dorée de la bibliothèque familiale pour les remplacer par mes cassettes dont j'organisais l'alignement et le classement comme preuve ultime de leur légitimité.

Ainsi, qu'une collection de cassettes de paroles de la vie quotidienne d'Orléanais enregistrées puisse être un objet scientifique digne d'intérêt pour un professeur d'université, agrégé et normalien, me permettait indéniablement de lier une histoire familiale dédiée à la valeur du capital scolaire et culturel avec une pointe d'originalité et de rébellion chez un étudiant qui cherchait sa place au sein d'une famille composée d'une mère professeure de lettres classiques et de huit enseignants (instituteurs et professeurs de collège et de lycée) parmi les frères, sœurs et leurs conjoints. En outre, le croisement de l'histoire familiale et de la politique française ne pouvait que faciliter cette position : avoir un an en 1968 puis quatorze en 1981 n'est pas anodin.

Si le corpus m'intéressait sous la forme d'un fonds, il n'en fut pas de même pour le contenu de la cassette et mon mémoire de DEA s'orienta vers un autre domaine : la politique et la presse écrite. Pourtant comme toute histoire marquante, celle-ci ne s'arrêta pas là et une dizaine d'année plus tard, en retrouvant l'université d'Orléans avec une thèse en poche à une époque bouleversée par les opérations de numérisation des fonds culturels et scientifiques, la collection de documents sonores du Corpus d'Orléans devenait plus désirable encore.

Parallèlement un parcours scientifique personnel m'orientait fortement vers l'enquête linguistique et la sociolinguistique de corpus.

2.2. Absence de données disponibles sur le français parlé [retour]

L'aventure d'ESLO commence par un constat cruel pour la linguistique de l'oral en France. Il n'existe pas de documents sonores disponibles pour un projet qui allie la description du français parlé et une perspective de didactique du français langue étrangère. Nous ne reviendrons pas sur l'absence de légitimité du français parlé dans la linguistique française. Celle-ci est présentée dans les travaux de Bergounioux (Bergounioux 1992[16]) et de Claire Blanche-Benveniste (Blanche-Benveniste & Jeanjean 1987[17]) :

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

Pourtant le vingtième siècle, berceau de la linguistique moderne, commençait par une avancée technologique qui aurait dû révolutionner très rapidement la linguistique. De fait, on pouvait estimer qu'avant 1900 l'impossibilité de rompre ou même tout simplement de capter le flux linéaire et éphémère des productions orales de la vie quotidienne ne permettait de décrire sérieusement ni les pratiques linguistiques ni même la langue comme objet d'une science. Le minimum requis pour atteindre un degré de scientificité suffisant était de passer par un formalisme qui reposait en premier lieu sur l'écriture de phénomènes stabilisés et normalisés.

On peut ainsi constater que l'absence de corpus coïncide avec l'absence d'enquêtes statistiques sur les langues parlées et les pratiques linguistiques en France. Ce parallélisme est peu relevé dans la littérature, pourtant il est bien réel et il résulte d'une même surdité de la part d'une science qui s'est systématiquement éloignée de l'observation des paroles entendues pour se consacrer à un objet formalisé et désocialisé : ne pas entendre la parole, c'est aussi ne pas entendre le locuteur et sa communauté linguistique dans la vie sociale.

Les premières enquêtes sur les langues de France ont été réalisées à la fin du XVIII^{e} et au début du XIX^{e} siècle. Parmi les rares enquêtes (Héran 1999), on peut en présenter trois :

- « *Le Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française* est un rapport rédigé par l'abbé Grégoire et présenté à la Convention nationale le 4 juin 1794 sur l'état de la langue française en France. Il s'appuie sur une véritable enquête sociolinguistique (1790), les réponses à un questionnaire de pas moins de quarante-trois questions relatives aux aspects internes et externes de la variété de langue parlée localement et aux mœurs et coutumes de la population. » L'abbé Grégoire avait

[16] BERGOUNIOUX, G. (1992). *Enquêtes, corpus et témoins en France, hier et aujourd'hui*.
[17] BLANCHE-BENVENISTE, C., & JEANJEAN, C. (1987). *Le français parlé: transcription et édition*.

adressé son questionnaire à tout un ensemble d'informateurs sur l'ensemble du territoire. Il en est résulté le constat qu'à peine un Français sur cinq a une connaissance précise de la langue française. On observe donc une grande diversité de langues (« patois »).

- L'enquête de Coquebert de Montbret en 1806 : Lors du recensement de 1806, Coquebert de Montbret lance une grande enquête sur les langues parlées par les Français. « Menée de 1806 à 1812, l'enquête Coquebert de Montbret produit une importante documentation sur l'ensemble des idiomes parlés dans l'Empire. Cette enquête demandait aux préfets une traduction littérale de la parabole de l'Enfant Prodigue dans le ou les « langages populaires » du département, ainsi que tout renseignement sur ces parlers. Les préfets étaient également invités à transmettre au ministère de l'Intérieur des textes, en vers ou en prose, utilisant ces idiomes. »

- L'enquête de Victor Duruy en 1864 : « Victor Duruy, ministre de l'Instruction publique sous le Second Empire, lance en 1864 une enquête statistique afin de mieux connaître la situation de l'enseignement primaire en France. Elle se démarque des précédentes par le degré de précision du questionnaire, destiné à dresser un tableau très précis de la situation de l'enseignement primaire en France. Pour la première et la seule fois dans l'histoire de la statistique scolaire en France, le questionnaire comprend une rubrique sur les "idiomes et patois en usage", à remplir par les inspecteurs primaires, les inspecteurs d'académie et les recteurs. » Les directives données en font une enquête statistique particulièrement précise et développée et surtout l'enquête Duruy apporte, jusqu'à l'enquête Famille de 1999, la matière documentaire la plus précise sur la situation sociolinguistique de la France.

Le but de ces enquêtes était donc de mieux connaitre la diversité linguistique en France, non pas afin de la promouvoir mais plutôt afin de « l'anéantir », pour reprendre les termes de l'intitulé de l'enquête de l'abbé Grégoire (1790-1794) : *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française*. On peut également utiliser l'exemple de l'enquête de Victor Duruy (1864). La rubrique sur les « idiomes et patois en usage » comprenait les questions suivantes :

« Existe-t-il des écoles où l'enseignement est encore donné en patois exclusivement ou en partie ? Nombre des écoles où l'enseignement est donné en totalité en patois ? En partie seulement ? Combien d'enfants savent le parler sans pouvoir l'écrire ? Quelles sont les causes qui s'opposent à une prompte réforme de cet état de choses ? Quels sont les moyens à employer pour le faire cesser ? »

Ces enquêtes ont permis de constater la grande diversité linguistique de la France avec précision, mais toujours avec le souci de substituer à cette diversité une unité, celle de la langue française. Mais surtout, comme le souligne Héran (Héran, 2002[18]), ces enquêtes ont toujours donné lieu à des analyses mal étayées. Les chiffres exploités par la suite sont de larges extrapolations qui sont systématiquement utilisées avec un objectif politique. Ces

[18] HERAN, F. (2002). « Les langues et la statistique publique, des comptages du Second Empire au volet linguistique de l'enquête famille ».

enquêtes sont loin de fournir une observation réelle des langues parlées et encore moins des productions des locuteurs :

> « *A l'autre bout du spectre il n'y aurait que 3 millions de Français capables de parler le français. Chiffres vraisemblables ? Chiffres invérifiables surtout. Comme le commente fort bien Anthony Lodge, les calculs de Grégoire visent d'abord à produire un effet politique. Dès cette époque, le chiffre est destiné à « faire nombre » : il fait sérieux sans qu'il soit nécessaire de le construire avec sérieux. En définitive le bilan de deux siècles de statistique publique en matière de langues est proche du néant*. » (Héran, 2002 :54)

Plus récemment les enquêtes se multiplient. Voici un tableau réalisé par joséphine Pasco dans le cadre d'un rapport de l'Observatoire des pratiques linguistiques (Baude, Pasco, Alessio, inédit) qui présente sept enquêtes de la statistique publique comportant des questions sur les langues, réalisées entre 1992 et 2013[19]. *[Observatoire des pratiques linguistiques]*

NOM-DATE	*NOMBRE DE PERSONNES INTERROGEES*	*RAPPORT AUX LANGUES*	*REMARQUES (Alexandra Filhon)*[20]
Enquête Éducation 1992	5 300 parents d'enfants scolarisés	Les interroger sur la principale langue utilisée dans le foyer.	Centrée sur la sphère familiale.
Enquête MGIS 1992 (Mobilité géographique et insertion sociale)	10 000 immigrés et descendants d'immigrés	S'intéressait aux langues d'immigration. Prévoyait la possibilité de citer deux langues d'enfance.	Centrée sur la sphère familiale.
Enquête Famille 1999 INSEE – INED	380 000 adultes	Possibilité de nommer deux langues.	Échantillon très vaste. Problème : ne dit rien sur les pratiques, privilégie les représentations et identités.
Enquête Histoire de vie 2003	8 403 adultes	Sur la construction des identités, dont quelques questions sur les langues.	Questionnaire plus étoffé que l'enquête Famille (mais échantillon plus limité) ; a permis de lier représentations et pratiques sociales et culturelles.

[19]Les questionnaires de ces enquêtes se trouvent en annexe.

[20]« Exposé introductif : les langues de l'immigration dans les enquêtes publiques » d'Alexandra Filhon, *Migrer d'une langue à l'autre ?*, Actes de la journée d'étude du mercredi 26 novembre 2014.

Enquête Information et Vie quotidienne 2004 et 2011 **INSEE et INED** (ont repris les questions de l'enquête Famille pour la rubrique sur les langues)	**2004** : Auprès de 10 400 ménages en France et concerne les individus âgés de 18 à 65 ans. **2011** : auprès de 14 000 personnes de 16 à 65 ans résidant en France métropolitaine.	Sont mesurées les compétences des adultes vivant en France en compréhension orale et écrite du français et en calcul. Questions sur la pratique des langues (y compris régionales) et sur leur transmission. Plusieurs réponses possibles.	
Enquête Trajectoires et Origines (TEO) 2008 **INSEE-INED**	22 000 personnes migrantes et issues de l'immigration	Le questionnaire explore l'histoire migratoire des personnes ou de leurs parents. Il étudie aussi la transmission des langues et de la religion dans le cadre familial. Enfin, il examine l'accès des individus aux ressources de la vie sociale (travail, logement, services, soins...) ainsi que les discriminations pouvant y faire obstacle.	S'intéresse plus aux locuteurs-personnes migrantes qu'aux langues.
ELIPA 2010-2013 **DGEF Ministère de l'Intérieur**	- 6 000 migrants - 3 interrogations sur 3 ans	Thèmes abordés : motif de la migration, projet migratoire, connaissance du parcours d'intégration dans 4 dimensions (acquisition de la langue, intégration personnelle et professionnelle, accès au logement et vie sociale), la connaissance du parcours administratif et de ses difficultés et l'évaluation du dispositif d'accueil.	A repris une partie de l'enquête IVQ. 97 langues ont été identifiées comme langues d'apprentissage de l'écrit.

On le constate, depuis une vingtaine d'années, les enquêtes croissent au moment même où la linguistique de corpus se développe. Là aussi l'observation des pratiques marche de pair avec l'observation de la langue parlée, même lorsque les enquêtes ne concernent pas l'oral. La similitude de traitement « de la parole » et des « pratiques » est alors saisissante.

Pourtant, comme nous l'indiquions, la découverte des technologies d'enregistrement était censée bouleverser cette conception de la langue et par-delà, de la linguistique. Ainsi quand Ferdinand Brunot installe les Archives de la parole (citation supra), il donnait toute l'ambition du programme scientifique dans son discours inaugural et le programme qu'il ambitionnait de mettre en place a été exécuté dans une perspective sociolinguistique :

> « *Par une anticipation remarquable, au moment même où Meillet soulignait que "Les linguistes inévitablement dominés par l'écriture, ne sauraient réfléchir assez à la façon dont toute notation trompe, de par sa nature même", Brunot pressentait que l'archivage non graphique du langage allait modifier considérablement les données dont pourraient disposer les linguistes : "on pourra un jour observer scientifiquement les manières si opposées dont dix personnes peuvent comprendre et décrire*

> *une phrase ou un vers selon leur origine, leurs habitudes, leur éducation générale et particulière, leur profession, leur disponibilité du moment, leur humeur, bref suivant une foule de conditions variables presque jusqu'à l'infini et où il importe cependant de démêler le fait accidentel et individuel d'avec le fait général et permanent*". (Encrevé 1988 :13[21])

Cette rupture entre linguistique de l'oral et linguistique de l'écrit recoupe une autre dichotomie au cœur du débat sur la linguistique, porté par Pierre Encrevé :

> « *Je voudrais conclure en relevant que si l'écriture a ouvert l'âge de la linguistique pour six millénaires, il me semble que depuis trente ans, avec la grande diffusion des moyens d'enregistrement magnétiques du son, quelque chose a changé parce que l'écriture n'enregistrait que l'invariant et que la bande magnétique enregistre toute la variation. Maintenant les linguistes peuvent disposer de données objectives sur la variation : « verba manent » et la totalité de l'information contenue dans la parole (âge, sexe, origine…) demeure aussi. Ce changement technique contribue, à mon sens, à déplacer la frontière entre linguistique et sociologie. Il est raisonnable que l'on tente d'utiliser davantage la science de la société qu'élaborent les sociologues pour construire et traiter le matériel linguistique maintenant que ce matériel peut être recueilli et conservé sans que l'opération de la collecte n'implique d'apurer la langue de toutes ses caractéristiques sociales* ». (Encrevé 1983 :23[22])

> « *Quand on écrit une langue, ce qu'on recherche par-delà toutes les variations, c'est l'invariant. On dit, en latin, que les paroles s'envolent et que les écrits restent. Mais quand les scripta prennent la place des verba, quelque chose disparait : la voix d'un locuteur donné est inséparable d'un certain nombre de traits très précis (l'âge, le sexe, peut-être la fonction, l'origine géographique et/ou sociale). L'écriture des mots que prononce cette voix laissera s'envoler toutes ces caractéristiques sociales ; demeurera l'invariant* ». (Encrevé 1983:3[23])

On le pressent, l'enregistrement de la parole comme l'observation des pratiques réelles bousculent doublement un champ dominé par l'écrit tout autant que par la négation de la nature sociale de la langue.

C'est ainsi que la linguistique reste enfermée dans une surdité qu'elle ravive parfois de manière magistrale. Ainsi un programme d'une très grande ambition comme celui du

[21] ENCREVE, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*.
[22] ENCREVE, P. (1983). « La « liaison » entre la linguistique et la sociolinguistique 1/2 ».
[23] Ibid.

Français fondamental mené à la sortie de la guerre et qui était censé fournir le lexique du « français parlé populaire » à partir d'une enquête statistique constate la pénurie de données sur le français parlé :

> « (…) *nous avons pu utiliser plusieurs enregistrements phonographiques conservés au musée de la parole (…) Ferdinand Brunot (…) avait enregistré une conversation entre lui-même et un ouvrier tapissier. Un autre disque à double face contient une conversation avec un habitant du faubourg Saint-Antoine. (…) Deux autres retracent le boniment d'un camelot. C'est tout ce que nous fournissait le Musée de la Parole pour le français commun.* » (Gougenheim et al., 1956 :62[24])

Devant cette pénurie, peu de surprise affichée par les auteurs qui n'hésitèrent pas à faire de nouveaux enregistrements sans se donner toutefois la peine de les conserver :

> « *Nous ne nous souciions pas non plus de la conservation des disques. Nous profitions largement de la possibilité qu'offrent des disques en papier magnétisé d'être effacés et de servir ainsi à plusieurs enregistrements successifs. Il aurait été beaucoup trop couteux de conserver tous les enregistrements comme de bons esprits nous le suggéraient.* » (Gougenheim et al., 1956 :63[25])

Le reste des informations méthodologiques sont tout aussi stupéfiantes et révèlent toutes l'incapacité de la linguistique à concevoir les productions orales comme le véritable objet de cette science. Cinquante ans après Brunot et son programme d'archives de la parole, la linguistique est plus que jamais éloignée de celui-ci comme le souligne Encrevé :

> « *Nous sommes entrés dans la légende !" écrivait Bruneau à Bruno. Ils ne tardèrent pas à comprendre qu'à poursuivre ce chemin ils n'entreraient que dans l'oubli, et sortiraient de la linguistique* ». (Encrevé 1988 :13)

Il faudra attendre dans un premier temps 1979 pour qu'une convention entre la BnF et le CNRS relance, dans le cadre des Atlas linguistiques, un projet minimaliste sous la forme d'une convention :

> *Le Greco n° 9 : Atlas linguistique de la France par région, du CNRS et le Département de la Phonothèque et de l'Audiovisuel de la Bibliothèque nationale signaient une convention qui avait pour objectif la conservation de tout ou partie des enregistrements sonores réalisés dans le cadre des*

[24] GOUGENHEIM, G. (1956). *L'élaboration du français élémentaire: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*.
[25] Ibid.

> Atlas linguistiques de la France*, puis des ethnotextes, ainsi que la constitution d'archives de sécurité.*
>
> *Le principe de cette convention reposait sur un système d'échanges réciproques : le chercheur décidant des enregistrements qu'il souhaitait voir conservés, les versait à la Bibliothèque nationale, accompagnés de leurs fiches descriptives. La Bibliothèque nationale en effectuait une copie, restituant les originaux au chercheur, ainsi qu'une ou deux bandes magnétique vierges pour ses travaux à venir. Par ailleurs, une copie des supports sonores archivés à la Bibliothèque nationale devait être déposée dans une institution régionale pour en assurer la consultation auprès du public en région.*
>
> *Cette convention avait donc pour objectif de permettre à un public de chercheurs d'avoir accès à un fonds patrimonial sonore normalisé et répertorié* [Cordereix 2002]

Les années qui suivirent continuèrent dans cette voie. Le français parlé, les langues parlées en France, les pratiques linguistiques quotidiennes restaient des domaines tenus à l'écart de la linguistique moderne. Le troisième millénaire ouvre néanmoins des perspectives autour de la place des corpus dans les pratiques scientifiques des humanités numériques.

Toutefois derrière la question des archives sonores et des pratiques linguistiques se cache une véritable perspective épistémologique pour la linguistique. Une présentation des effets de cette posture de la linguistique française sur le projet ESLO est présente dans Bergounioux & Baude *(Bergounioux&Baude,2015 :6)*[26] et dans différentes communications sur ESLO :

> Bergounioux & Baude *(Bergounioux & Baude,2015:6,* [27] *« 2.Eslo, un enquête en son temps : enjeux, méthodes résultats »)*
>
> *« Dans l'attention accordée aux parlures populaires, argots et patois, le français ordinaire, ravalé par la norme littéraire inculquée à l'école, est d'emblée écarté. Gilliéron et Edmont laissent en blanc sur leur atlas Paris et sa région et F. Brunot, au moment de la création des Archives de la parole, entreprend un inventaire dans les Ardennes, reprenant l'étude dialectale de Charles Bruneau (1913), puis se dirige*

[26] BERGOUNIOUX, G., & BAUDE, O. (2015). « ESLO, UNE ENQUÊTE EN SON TEMPS : ENJEUX, MÉTHODES ET RÉSULTATS ».
[27] Ibid.

ensuite vers le Centre et le Limousin, sans graver sur les cylindres Pathé des échanges quotidiens en français à l'exception d'un très court dialogue avec un menuisier (Cordereix, 2006).

La demande sociale aurait pu provenir de l'enseignement du français langue étrangère en un temps où les voyages et les séjours linguistiques ne facilitaient pas la connaissance des langues. Les cours phonographiques semblent n'avoir pas produit plus de résultats que les cours radiodiffusés. La demande concernait préférentiellement l'usage d'un registre très soutenu, très artificiel aussi, comme on l'entend dans les enregistrements de pièces de théâtre ou d'émissions de la TSF avant 1960.

L'intérêt pour le français ordinaire emprunte une voix frayée aux Etats-Unis trente ans auparavant pour l'établissement des fréquences verbales. La collecte du Français Fondamental, motivée par le succès du BASIC English, centre son attention sur le lexique mais innove en abandonnant le dépouillement des textes à quoi sont substitués des enregistrements commandés pour l'occasion. Réalisée par le CREDIF au cours des années 50 pour établir la liste des mots les plus courants dans des échanges quotidiens (Gougenheim et al., 1956), l'enquête témoigne rétrospectivement d'une certaine naïveté sociologique et la technique de capture du son a été, pour des raisons économiques, très sommaire. Comme l'objectif était de constituer un vocabulaire minimal, à la rigueur une syntaxe élémentaire, aucune attention n'a été portée aux données sonores et les enregistrements, effectués sur un support de piètre qualité, ont été effacés au fur et à mesure de l'exécution de transcriptions qui ne s'embarrassaient pas de relever des variations. La modestie des consignes données contraste avec l'importance des recommandations que l'équipe d'Aix-en-Provence prodiguera (Claire Blanche- Benveniste, 1987). L'ampleur des moyens engagés, en dépit de ses prolongements dans la diffusion du FLE, aura peu de répercussions sur la description du français vivant et sur les méthodes des linguistes en France, passé une polémique vite éteinte (Cohen, 1955).

Tel est le constat de carence que dressent les universitaires anglais qui, à la fin des années 60, se proposent de confectionner un manuel du français à l'usage de l'enseignement secondaire en se fondant sur des enregistrements. Ils ont tôt fait de constater qu'il leur était impossible de se procurer les matériaux dont ils avaient besoin dans

> *les fonds d'archives alors que leur préoccupation n'était plus l'établissement de statistiques lexicales mais l'accès à la diversité des réalisations du français parlé auxquelles seront exposés les lycéens britanniques. Au nombre des dissymétries entre l'enquête du Français Fondamental et ESLO, on peut relever l'inégale légitimité des acteurs : un parrainage par des linguistes parisiens reconnus (A. Sauvageot, G. Gougenheim) avec le soutien de l'Education nationale d'un côté, des professeurs de langue étrangère issus de facultés qui ne sont pas les plus prestigieuses.»*

En effet, si le *Français fondamental* était motivé, dès ses débuts, par le constat de la pauvreté des enregistrements de français parlé disponibles, lorsque les auteurs du projet entreprirent eux-mêmes de recueillir des données, ils n'hésitèrent pas à effacer les enregistrements audio dès que ceux-ci étaient transcrits. Derrière la décision de ne conserver que le lexique à des fins statistiques transparaît la non légitimité de la syntaxe du français ordinaire (Abouda & Baude, 2005 [28]).

La première caractéristique du projet de l'ESLO est donc d'apporter une reconnaissance et une légitimité au français parlé en collectant et en rendant disponible un corpus de données authentiques.

> (Abouda & Baude, 2005 [29], « Du français Fondamental aux ESLO »)
>
> *La conservation et la possibilité de réutilisation des matériaux recueillis étaient dès le départ considérées par les initiateurs d'ESLO comme deux objectifs fondamentaux. Ce choix s'est concrétisé de différentes manières.*
>
> *1. D'abord par le catalogage et l'indexation : l'équipe de l'ESLO a publié en 1974 un catalogue descriptif et analytique qui répertoriait les enregistrements avec : résumé du contenu, indexation des questions, organisation du questionnaire, catégorisation sociologique précise des locuteurs et description de la situation d'enquête.*
>
> *2. La conservation des données primaires (enregistrements et documents d'enquête).*
>
> *3. Les transcriptions. Bien qu'une transcription intégrale fût difficilement envisageable pour un corpus estimé à plus de 4 millions de mots, l'équipe a entrepris immédiatement la transcription*

[28] ABOUDA, L., & BAUDE, O. (2005). « Du Français Fondamental aux ESLO ».
[29] Ibid.

> *d'extraits qui se voulaient représentatifs et qui recouvraient toutes les catégories des témoins (INSEE et AM).*
>
> *4. La diffusion du corpus*
>
> *Outre l'annonce systématique dans les articles de la disponibilité du corpus, le catalogue précise dès la page 4 :*
>
> *« Les transcriptions et enregistrements sont disponibles à tout chercheur intéressé, contre remboursement des frais de matériaux et de copiage ; (…) Des listes de transcriptions et enregistrements sont disponibles à ceux qui s'adressent à nous. » (Lonergan et al., 1974:4).*

Il s'agit d'ailleurs d'un des objectifs du projet clairement affiché parmi la liste des six déclarés dans la présentation du catalogue de 1974.

2.3 Légitimité d'une sociolinguistique sur corpus [malentendus] ESLO [retour]

La place des corpus en linguistique donne lieu à des polémiques constantes. Laks (2008) souligne les relations anciennes entre la linguistique et les corpus :

> « *En linguistique, il n'en a pas toujours été ainsi. Aux 19e et 20e siècles, les sciences du langage et leurs précurseurs se sont constituées comme des sciences du datum, inscrivant leur démarche dans la dynamique épistémologique qui depuis la Renaissance faisait émerger la science moderne comme une systématique adossée à de larges compendiums de faits. Du point de vue historique, la notion de corpus apparait en effet comme très ancienne, mais elle joue un rôle de première importance dans le développement de la pensée scientifique moderne. (…) C'est cette approche que je retiens ici : la linguistique de corpus, loin de constituer un courant nouveau, apparaît en science du langage comme une orientation ancienne voire très ancienne, même s'il est vrai qu'elle ne fait retour en pleine lumière que nimbée de l'aura de technologies et d'outils très sophistiqués* ».(Laks 2008 :5)[30]

Au XXIe siècle nous ne pouvons que constater une ambigüité forte sur la notion de corpus, comme le prouvent ces différentes citations

[30] LAKS, B. (2008). « Pour une phonologie de corpus ».

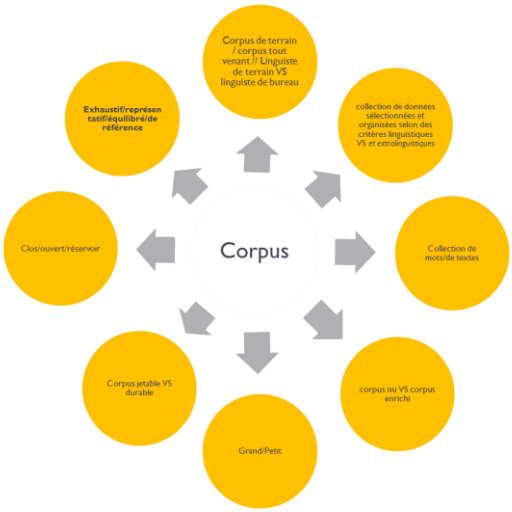
« *La francophonie cède à l'engouement pour les corpus, avec quelque retard par rapport aux initiatives et aux recherches anglo-saxonnes. Les rencontres et les projets s'enchaînent, non sans quelque confusion : le mot corpus est tiraillé dans des directions parfois bien éloignées. La réalité même des corpus a en outre beaucoup évolué. La vieille question de la représentativité des corpus ressurgit. Il importe d'évaluer si les termes mêmes dans lesquels elle se posait se sont ou non déplacés* ». (Habert, 2001:8).

« La notion de corpus paraît, de prime abord, assez simple et bien ancrée dans certaines traditions des sciences humaines et sociales, philologique ou juridique par exemple. Il s'agit d'un recueil formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline. Néanmoins, dans le champ linguistique, la notion s'est complexifiée au cours des dernières décennies en fonction de la diversité des pratiques et des objectifs assignés à la constitution et l'exploitation des corpus». (Mellet, 2002 :69).

« Le corpus –la notion et l'objet– risque d'être victime aujourd'hui en France de son succès. Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence ; plus un linguiste, surtout, qui ne le manipule, le caresse ou le maltraite » (Mayaffre, 2005 :16).

« *Deux conceptions du corpus : sac de mots ou archive de textes* ? » (Rastier, 2005 :81).

Nous pouvons déterminer les axes de tension qui conditionnent l'acception de cette notion :



Le premier axe est celui qui oppose la linguistique de terrain et la linguistique de bureau :

« Pourtant il y a encore des linguistes pour prendre des corpus comme du tout-venant discursif, et se lancer sur des corpus provenant de terrains dont ils n'ont pas grande connaissance, voire aucune » [Cappeau & Gadet 2007]

Le second axe oppose les collections sélectionnées et organisées selon des critères linguistiques à celles qui utilisent des critères extralinguistiques :

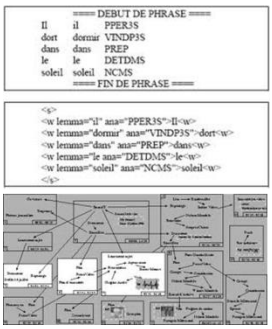
« La linguistique de corpus insiste sur le caractère restrictif du corpus : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue » [Habert 2000].

Le troisième axe oppose la collection des listes de mots à celle de textes structurés :

« Cependant, un corpus n'est pas plus un sac de mots qu'un nébuleux intertexte. Il est structuré d'une part en fonction d'une typologie des textes, qui se reflète dans leur codage, et d'autre part, dans chaque utilisation, par des sélections raisonnées de sous-corpus « [Rastier 2005].

Le quatrième axe oppose des corpus nus à des corpus enrichis :

« Il dort dans le soleil » VS



Le cinquième axe oppose les objectifs de taille posés comme critère premier :
"Gros c'est beau" (la Toile) VS "ensembles aux conditions de production et de réception plus nettement définies et corrélées à leurs caractéristiques langagières" [Habert 2001]
"A partir de quand un corpus oral est-il grand ? (…) quand il dépasse 1 000 000 de mots (2000 pages) ? Ces rappels chiffrés donnent la mesure du caractère à la fois apparemment vaste et si minime (quant aux chances de documenter certains phénomènes, ou d'identifier des contraintes) des données recueillies". [Cappeau & Gadet 2007].

Le sixième axe oppose des corpus strictement conçus pour une étude spécifique à des corpus réutilisables.

Le septième axe définit le périmètre du corpus :

- clos : constitué une fois pour toute, avec comme objectif de présenter un corpus "complet" [Full-text Corpus, Kennedy 1998].

- ouvert : prévu pour incorporer de nouvelles données par la suite, ou en continu, exemple la Toile comme corpus dynamique [Renouf 2002]

- réservoir : banque de donnée permettant l'élaboration d'un corpus par extraction (exemple BNC) [Habert 2001, 2006]

Enfin le huitième axe oppose les corpus exhaustifs aux corpus représentatifs :

- exhaustif (cf. corpus clos, contient toutes les données relatives à l'étude, ex les œuvres d'un auteur)
- représentatif : notion vague [Habert 2001]
  - représentatif des genres [Biber 1993]
  - représentatif d'un échantillonnage sociologique,
  - représentatif d'une situation de production, etc.
- équilibré (panaché) : représentativité des genres (ex ICE) ou d'une langue (corpus de référence).
- de référence :

A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. The model for selection usually defines a number of parameters that provide for the inclusion of as many

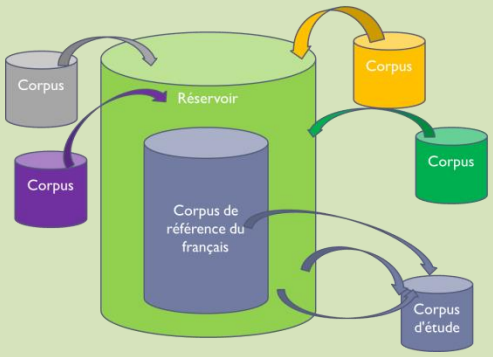
> *sociolinguistic variables as possible and prescribes the proportions of each text type that are selected. A large reference corpus may have a hierarchically ordered structure of components and subcorpora (Sinclair 1996:324).*

Ces axes permettent de définir un corpus qu'on ne saurait considérer comme un « fourre-tout » dans lequel on pioche des données à des fins d'analyses (Laks 2008). Au-delà d'un ensemble de données le corpus reste toujours à construire :

> *Bases et banques de données sont des ensembles de données, souvent composites, agrégées sans autre objectif explicité que quantitatif et de documentation empirique. Les bases de données sonores modernes comprennent ainsi des éléments, de types et de statuts très divers, voire hétéroclites (émissions radiodiffusées ou télévisées, interviews, discours publics, enregistrement ethnographiques, enquêtes phonologiques etc.). Comme le note Blanche-Benveniste (2004), elles apparaissent dès le début du 20e siècle avec les premières techniques d'enregistrement phonographique et se multiplient très récemment avec le développement des archives sonores. Elles sont aujourd'hui potentiellement surabondantes, ne nécessitant qu'un travail de collecte et d'établissement des métadonnées (Cf. la mise en ligne des archives de l'INA). Enfin, banques et bases de données sonores sont peu, ou pas, tributaires d'hypothèses linguistiques et phonologiques précises et ne peuvent servir qu'à des fins documentaires externes à l'analyse phonologique proprement dite. Lorsque l'on dispose d'une base de données sonores, le corpus correspondant reste à construire.* (Laks 2008 :9[31])

Ainsi un corpus de « référence » ne peut être conçu comme un simple entassement de données et il ne peut être non plus une base partageable pour différentes analyses sans un réel travail réflexif. C'est le sens du schéma suivant qui résume cette conception dans le cadre d'un vaste projet institutionnel de constitution d'un corpus de référence du français (Baude 2012-2013) :

[31] Ibid.



[réservoir]

Cependant cette architecture n'est possible que si les corpus produits par les différents projets, laboratoires et chercheurs sont à la fois réalisés dans une perspective scientifique maitrisée et documentée, et formatés afin de favoriser l'interopérabilité sans dissimuler derrière une uniformité de façade des approches théoriques diverses.

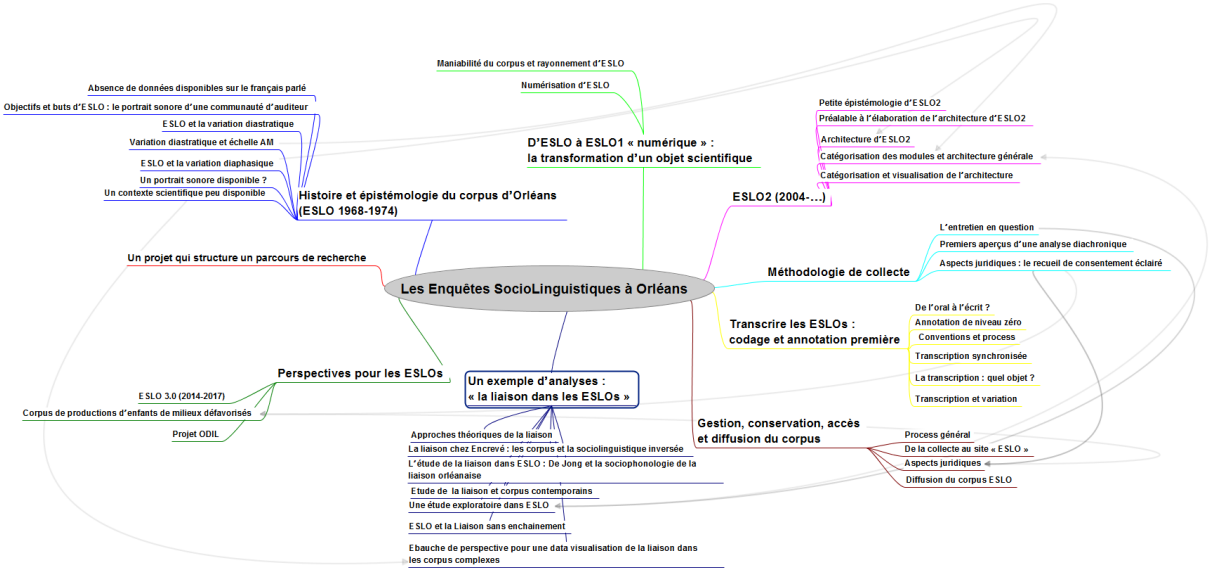
In fine c'est encore une fois l'objet même de la linguistique qui se définit ici :

> « *Comme je l'ai montré ailleurs (Laks (2011b), c'était déjà la position de Saussure qui défendait le primat de la linguistique de la parole comme condition sine qua non d'une grammaire de la langue. Thématisant ainsi les pratiques langagières dans leur contexte écologique social et culturel, les linguistiques de l'usage redonnent un statut central à la description linguistique, aux analyses distributionnelles, statistiques et fréquentielles, et se situent donc clairement dans la mouvance des linguistiques de corpus. Ce sont des linguistiques du datum et corrélativement, la variation et l'hétérogénéité internes y sont reconnues. Mais surtout, dans les modèles basés sur l'usage, cette variation sociolinguistique, tant synchronique que diachronique, retrouve un statut systémique et fonctionnel. Enfin, tant du point de vue de l'instanciation cognitive des régularités linguistiques que du point de vue de l'apprentissage situé de la langue et de la grammaire, ces approches sont parallèles aux approches connexionnistes et neuromimétiques et plus généralement ressortissent aux modélisations dynamiques du langage. Avec toutes leurs variantes et approches connexes, grammaires de construction, grammaires exemplaristes et occurentialistes, grammaires discursives, neurales et cognitives, grammaires stochastiques et probabilistiques etc., les modèles basés sur l'usage et les linguistiques du datum ont profondément modifié le champ linguistique international et, marginalisant la linguistique cartésienne, constituent le paradigme dominant de ce début de 21^e^ siècle*. » (Laks 2010 :19) [32]

[32] LAKS, B. (2010). « Langage et variation : pourquoi y a-t-il de la variation plutôt que rien ? ».

Ainsi, au moment où les linguistes s'engouffrent dans les humanités numériques, et plus que jamais si c'est bien le point de vue qui crée l'objet, c'est la notion même de corpus qui entre en jeu dans une démarche d'adéquation observationnelle.

3. Les Enquêtes sociolinguistiques à Orléans (ESLO) [retour]



3.1 Histoire et épistémologie du corpus d'Orléans (ESLO 1968-1974) [retour]

L'enquête sociolinguistique à Orléans (ESLO) est un grand projet réalisé par une équipe franco-britannique à la fin des années 1960 dans le but de collecter un corpus de français parlé pour des analyses sociolinguistiques et des applications didactiques. L'entreprise, qui dura plus de cinq ans entre la conception et la réalisation, donna naissance à l'un des corpus les plus vastes de français oral, le Corpus d'Orléans : 350 bandes magnétiques représentant quelques 317 heures d'enregistrements, ce qui correspond à quelques ± 4 500 000 mots. Ce chapitre présente les objectifs, la méthodologie de cette enquête et la resitue dans un cadre épistémologique où la sociolinguistique et la linguistique de corpus naissaient.

Articles et livre :	
	○ 2015, ESLO, une enquête en son temps : enjeux, méthodes et résultats, https://halshs.archives-ouvertes.fr/halshs-01165934
Communications orales :	
	○ 2006, Constitution et exploitation d'un grand corpus de "données situées" Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), https://halshs.archives-ouvertes.fr/halshs-01165954 ○ 2007, Le corpus d'Orléans, https://halshs.archives-ouvertes.fr/halshs-01166003 ○ 2008, Du Français Fondamental aux ESLO, https://halshs.archives-ouvertes.fr/halshs-01162533
Documents :	
	○ L'enquête sociolinguistique sur le français parlé à Orléans, http://www.nakala.fr/data/11280/97b0c99a

- Catalogue ESLO http://www.nakala.fr/data/11280/deecd9b4
- Présentation projet scientifique : http://eslo.huma-num.fr/index.php/pagepresentation/pageprojetscientifique
- Affiche Eslo2 (metadata)
- Bulletin de participation ESLO2 (metadata)
- Complément au guide du transcripteur: Lexique (metadata)
- Formulaire de consentement ESLO2 (metadata)
- Formulaire locuteur ESLO2 (metadata)
- Guide du matériel ESLO2 (metadata)
- Guide du transcripteur/relecteur des ESLOs V1 (metadata)
- Guide du transcripteur/relecteur des ESLOs V2 (metadata)
- Guide du transcripteur/relecteur des ESLOs V3 (metadata)
- Guide du transcripteur/relecteur des ESLOs V4 (metadata)
- Plaquette de présentation ESLO2 (metadata)
- Procédure dépôt ESLO2 (metadata)

3.1.1 Objectifs et buts d'ESLO : le portrait sonore d'une communauté d'auditeur [retour]

L'origine du corpus repose sur une perspective didactique de l'enseignement du français langue seconde en Grande Bretagne dans un contexte sociopolitique des années soixante :

> Bergounioux & Baude *(Bergounioux&Baude, 2015:6,* [33] *« 2.Eslo, un enquête en son temps : enjeux, méthodes résultats »)*
>
> L'Angleterre des années 60 est traversée par un courant de rénovation dans la vie politique porté par l'aile gauche du Labour Party qui, avec la fin de l'empire colonial, tourne ses regards vers l'Europe. Cette orientation a des répercussions sur le système d'enseignement. La scolarisation secondaire, dominée traditionnellement par le modèle des public schools, s'ouvre à tous et des allocations d'études sont proposées aux étudiants pour lutter contre la sélection sociale. Il s'ensuit une demande pour une pédagogie rénovée de l'apprentissage des langues qui, refusant le dilemme entre des applications immédiates au domaine commercial et l'acquisition d'un signe électif d'appartenance aux classes dominantes, devient un élément central de la culture transmise par le système éducatif. A l'opposé de la prédilection pour l'écrit et les auteurs classiques au principe des cours dispensés jusqu'alors, la connaissance d'au moins une langue étrangère à destination de publics moins sensibles au prestige de la tradition doit se faire avec des méthodes modernes et des contenus modernes selon le

Laks, "Pour Une Phonologie de Corpus."

[33] BERGOUNIOUX, G., & BAUDE, O. (2015). « ESLO, UNE ENQUÊTE EN SON TEMPS : ENJEUX, MÉTHODES ET RÉSULTATS ».

jugement de quelques enseignants du supérieur impliqués dans la formation des professeurs.

Cet objectif a un effet sur la définition de la langue collectée, ancrant le projet dans une recherche de données issues de l'observation des pratiques linguistiques de la vie quotidienne des locuteurs et dans une collecte systématique fondant un véritable corpus. ESLO est indéniablement le premier corpus en France qui cherche à atteindre un degré de représentativité « structurée » du français parlé :

> *" Les origines de l'ESLO remontent en 1966, à la période de la "révolution audio-visuelle" de l'enseignement des langues modernes en Grande Bretagne. L'introduction de nouvelles techniques et surtout, l'importance croissante accordée à la parole non littéraire, faisaient ressortir un besoin aigu d'échantillons authentiques de français parlé spontané. Mais 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., 1974:4).*

Cet objectif se traduira en une méthode de langue particulièrement innovante à la fois sur le contenu du corpus comme ressource pédagogique et sur la place des données authentiques dans la méthode de langue. On en retrouve toute la saveur dans cet extrait du manuel *Les orléanais ont la parole* :

19 La femme au travail — pour

Cette dame était autrefois comptable, comme Monsieur QC. Comme beaucoup d'autres femmes, elle a dû abandonner son travail, et comme au moins certaines autres femmes, elle en a quelques regrets.

QUESTIONS

1 Pourquoi cette dame aurait-elle aimé continuer à travailler?
2 Qu'est-ce qui l'en a empêchée?
3 Qu'est-ce qui compense dans une certaine mesure l'abandon de son travail?
4 Qu'est-ce qui a changé à Orléans depuis quelques années?
5 Que fera sa fille si elle se marie?

TRANSCRIPTION

Transcrivez :
1 le début du texte, jusqu'à : « ... plus seule finalement »
2 « enfin remarquez moi ... toute la journée hein »
Marquez sur votre transcription toutes les élisions et toutes les liaisons.

EXERCICES

a. Langue parlée, langue écrite

Regardez la dernière partie de la première section transcrite :
« parce que ... finalement ».
Écrivez une ou deux phrases qui expriment dans un style suivi tout ce que dit Madame KH ici.

b. Langue parlée

Consultez, à la fin de ce livre, la note dans le *Glossary* sur les locutions *enfin*.
Cherchez dans les deux sections transcrites 5 ou 6 expressions qui sont des locutions *enfin*. Indiquez quelles sont celles qui font appel à l'interviewer, et celles qui marquent des hésitations.

c. On

Réécrivez le texte pour faire une liste de verbes dont le sujet est « on ».
Indiquez dans une deuxième colonne les personnes représentées par ce pronom.

d. Phrases conditionnelles

Madame KH illustre un type de structure pour des phrases conditionnelles qui appartient seulement à la langue parlée.
1 *Structure habituelle (langue écrite ou parlée) :*
« je me serais épanouie davantage si j'étais restée de chez moi »
Complétez les phrases qui suivent sur ce modèle :
a J'aurais continué à travailler si je ...
b Je me serais sentie très seule si mes enfants ...
c J'aurais aimé faire des études supérieures si mes parents ...
d Je serais partie plus tôt si la circulation ...
e J'aurais voyagé à l'étranger si vous ...
f Je serais sûrement repartie travailler si ...
2 *Structure de la langue parlée seulement :*
« j'en aurais pas eu tant je serais sûrement repartie travailler »
Refaites les phrases *a - f* ci-dessus pour leur donner cette forme.
3 Cherchez à la fin du texte encore un exemple d'une phrase conditionnelle qui a une structure parlée semblable.

e. Ordre des mots

« elle est aide-comptable dans l'usine où travaille mon mari »
Complétez les phrases suivantes en y insérant un verbe convenable :
1 Nous allons passer les vacances à Limoges où ..... mes parents.
2 C'est la même profession que ..... le frère de mon mari.
3 Etes-vous au courant des événements dont ..... cet article ?
4 Il n'y avait pas de bruit dans la chambre où ..... les enfants.
5 Il a prononcé les mêmes paroles que ..... le Président de la République.

La suite des objectifs est en droite ligne de ce principe fondateur :

> *« Au cours de la planification de l'entreprise, il est devenu évident que la description linguistique aurait à tenir compte de l'identité sociale de*

chaque locuteur et de la situation de communication dans chaque cas, ce qui plaçait le projet solidement dans le camp de la discipline, alors naissante, de la sociolinguistique. L'ESLO avait donc de multiples buts :

- de réunir un corpus d'enregistrements du français parlé, pris à l'intérieur d'une société urbaine, le choix des témoins devant être gouverné par des critères sociologiques explicites afin d'assurer la représentativité du Corpus,

- de transcrire un échantillon représentatif du corpus,

- de préparer et de publier un catalogue descriptif et analytique des documents sonores et écrits afin de les rendre disponibles aux chercheurs,

- de créer des ensembles pédagogiques pilotes destinés à l'enseignement secondaire et supérieur,

- de réaliser des études pilotes de description et d'analyse linguistique (Lonergan et al., 1974:4)».

Ces objectifs se concrétisent autour de la notion de « portrait sonore d'une ville » :

> Bergounioux & Baude *(Bergounioux&Baude, 2015:6,* [34] *« 2.Eslo, un enquête en son temps : enjeux, méthodes résultats »)*
>
> *Le titre « Portrait sonore d'une ville » résume plusieurs intentions. La ville se situe à l'opposé de l'enquête dialectale sur un terroir par le choix d'une agglomération en pleine croissance, où le brassage des populations dilue la transmission endogène. Avec le son, on s'affranchit au moins partiellement de l'écrit, de l'effacement de la variation à quoi avaient abouti les transcriptions. Avec le« portrait », on renonce à mettre l'accent sur un état de langue pour concentrer l'attention sur les locuteurs, leurs interactions. Pourquoi avoir fait choix d'Orléans ? Les raisons ont été explicitées par les auteurs eux-mêmes. Il s'agissait de recenser une réalisation du français qui ne soit pas identifiée à un accent régional marqué, dans une ville d'une certaine importance qui venait de rouvrir son université, bien reliée à Paris et qui n'en soit pas trop distante pour des raisons logistiques. (…)*
>
> *Les enquêteurs ont concentré l'essentiel de leur collecte sur le premier semestre 1969, associant à leur démarche des assistants et des doctorants français et anglais. (…)*

[34] Ibid.

Si ESLO n'a pas joué un rôle déterminant dans les principes de la sociolinguistique, une révision venue plutôt des Etats-Unis avec W. Labov dont les publications fondatrices sont contemporaines (1966), l'ensemble constitue plus qu'un document d'une qualité scientifique exceptionnelle, une nouvelle pratique méthodologique, en linguistique de corpus.

Ainsi en quelques mois l'équipe franco-britannique va réaliser une vaste enquête sociolinguistique en collectant plusieurs centaines d'enregistrement dans une opération unique en son genre.


POUR SAISIR UNE IMAGE
(celle de la France et des Français)
QUINZE UNIVERSITAIRES BRITANNIQUES COMPTENT SUR LEUR MAGNÉTOPHONE... ET SUR 360 ORLÉANAIS
GRÈVE-SU
Le courant
matin penda
Certains industrie
360 interviews

La *République du centre*, Jeudi 3 avril 1969

Comment constituer ce « portrait sonore » représentatif d'une communauté linguistique ? Celui-ci se fera selon deux axes qu'on peut schématiser par une définition traditionnelle, même si, ici comme ailleurs, elle ne correspond pas véritablement à un cadre théorique efficient : la variation diastratique et la variation diaphasique. L'équipe s'appuiera sur une sociologie rigoureuse pour la première et tâtonnera à partir des prémices de la sociolinguistique interactionnelle pour la seconde.

3.1.2 ESLO et la variation diastratique [analyse] [retour]

Le premier geste de l'équipe d'ESLO pour constituer un corpus représentatif est celui d'une recherche classique d'un échantillonnage réalisé de la manière la plus rigoureuse qui soit. Ainsi la recherche d'une communauté d'auditeurs est appréhendée par un échantillonnage précis et une tentative de classification des agents selon des critères socioculturels (échelle AM). Cette tentative constitue un module du corpus :

- 157 entretiens / 182h30' (but constitution d'une gamme sociologiquement représentative de témoignages à contenu constant).

http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000095934
http://cocoon.huma-num.fr/exist/crdo/meta/crdo-COLLECTION_ESLO1

> Bergounioux & Baude *(Bergounioux&Baude, 2015:6,* [35] *« 2.Eslo, un enquête en son temps : enjeux, méthodes résultats »)*
>
> *Les témoignages pour obtenir une photographie des différentes compétences langagières ont été sollicités par le biais de l'INSEE qui a fourni à l'équipe un échantillon de la population urbaine constitué par tirage au sort à partir de ses fichiers, combinant des critères d'âge, de sexe et de catégorie socio-professionnelle (CSP). La désignation aveugle des témoins aura pour effet une distorsion des réponses : l'inégale légitimité à se concevoir comme un représentant attitré de l'usage du français conduira la plupart des personnes sollicitées dans les milieux modestes à se récuser.*

Cette rigueur atteint-elle véritablement son but dans le cadre d'un corpus oral représentatif des pratiques linguistiques ? Rien n'est moins sûr et ceci doit nous interroger sur le risque d'une tentation scientiste, là où une nouvelle méthodologie est nécessaire. En effet les objectifs annoncés ne passeront pas le test des contraintes d'une collecte de terrain. Pour un panel recherché de 600 locuteurs, l'équipe se rabattra sur 147 entretiens réalisés. Visiblement la difficulté de trouver des orléanais qui acceptent de raconter leur vie devant un micro dans le cadre d'une enquête sociolinguistique avait été sous-estimée.

> *Baude & Dugua 2015*
>
> *Il était prévu un tirage au sort par les services de l'INSEE d'un échantillon de six cents témoins selon ces trois critères : sexe (deux catégories), âge (trois tranches) et catégorie socio-professionnelle (cinq catégories). Le croisement de ces critères équilibrés donnait trente sous-groupes de vingt témoins. Les difficultés rencontrées par les chercheurs ont considérablement affaibli cette démarche. Le taux de refus était beaucoup plus important dans certaines catégories sociales. Au terme de l'enquête, seul un quart du panel a pu être réalisé. (Baude & Dugua 2015:8)*

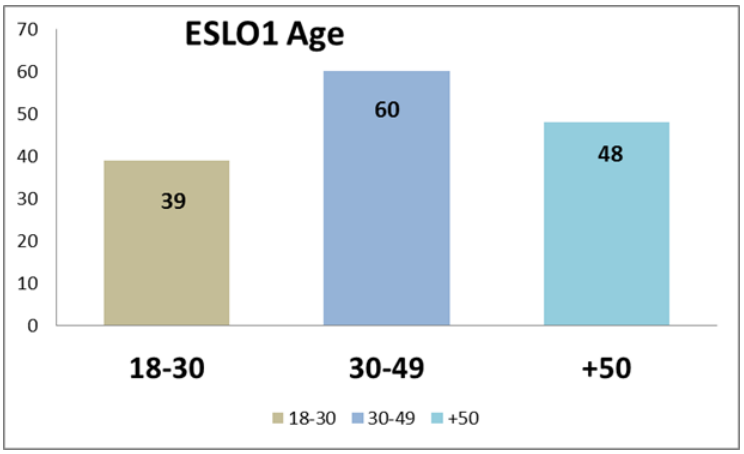
[35] Ibid.

Rien d'étonnant, les sociologues savent depuis longtemps que certaines catégories sociales ne s'autorisent jamais à s'exprimer (Beaud 1997).

Cette difficulté est riche d'enseignement concernant la méthodologie à élaborer et les effets des choix méthodologiques sur l'architecture du corpus et sa représentativité, enfin sur la « nature sociale » même de la langue si difficile à observer.

Ainsi si l'échantillonnage en sexe est respecté :

il n'en est pas de même pour la répartition en âge :



et encore moins en CSP :

	Patrons	Professions libérales	Cadres moyens	Employés	Ouvriers	Personnel de service	TOTAL
Prévus	100	100	100	100	100	100	600
Répartition sur 147	24	24	24	24	24	24	~147
Réalisés	27	27	16	28	42	7	147

Une analyse de la distorsion entre le corpus visé et le corpus réalisé est nécessaire afin de mettre en évidence ce qui pose une réelle question pour une science de l'observation : comment mesurer l'amplitude de la variation causée par la méthodologie ?

Cette étape est le plus souvent oubliée dans les analyses de corpus où l'on conserve l'objectif déclaré comme référence de la nature des données au détriment d'une analyse de leurs conditions de production.

Cette expérience d'ESLO aura une très forte incidence sur la méthodologie d'ESLO2.

3.1.3 Variation diastratique et échelle AM [corpus] [analyses] [retour]

La méthodologie d'échantillonnage d'ESLO1 est très révélatrice d'une période où la sociologie quantitative se fondait principalement sur une catégorisation des témoins selon les caractéristiques d'âge, de sexe et de profession. Cette méthodologie était recherchée à des fins de rigueur scientifique (Mullineaux & Blanc 1982) avec la réussite très mitigée décrite suppra.

Au-delà des critères de classifications classiques issus de la sociologie quantitative et reprise en partie dans le cadre de travaux en sociolinguistique à partir d'une définition d'une « stratification sociale » (Labov 1973), ESLO1 présente une tentative particulière de catégorisation menée par Alix Mullineaux (Mullineaux & Blanc 1982[36]). Aux critères précédemment cités, elle a ajouté, dans un premier temps, le niveau et l'âge de fin d'études puis elle a envisagé de compléter ces critères par une évaluation du capital culturel (repéré à l'issue des entretiens et notamment en prenant en compte les questions sur les goûts et pratiques culturelles) de chaque locuteur. L'objectif était alors de diviser le corpus en cinq groupes (échelle « AM » – du nom de l'auteur – de A à E).

> *« Eslo, un enquête en son temps : enjeux, méthodes résultats ») (Bergounioux & Baude, 2015:8,*[37]
>
> *Le développement le plus intéressant concerne le glissement dans les critères sociologiques où est substituée, de façon hésitante mais cohérente, aux cadres déterministes de Bernstein, une différenciation entre ce que P. Bourdieu, rapidement consulté, caractérisera ultérieurement comme la distinction que produit la composition relative du capital économique et du capital culturel dans la définition sociale d'un agent. L'échelle AM (pour Alix Mullineaux, une chercheuse associée au groupe dont les propositions figurent en appendice dans le Catalogue) pondère le classement INSEE en prenant en considération l'âge de fin d'étude,*

[36] MULLINEAUX, A., & BLANC. (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 ».
[37] BERGOUNIOUX, G., & BAUDE, O. (2015). « ESLO, UNE ENQUÊTE EN SON TEMPS : ENJEUX, MÉTHODES ET RÉSULTATS ».

> *correspondant dans le capital culturel à la part généralement déterminante du capital scolaire.*

Les perspectives esquissées par Alix Mullineaux sont très novatrices et, outre l'apport de cette réflexion sur les travaux de classification qui résonnent de façon particulière après des travaux en sociologie comme ceux de Desrosières (2008[38]), nous pouvons y discerner les enjeux d'une linguistique de corpus qui souhaite maitriser la relation entre les catégories « pré et post corpus ». En effet la véritable échelle AM aurait existé seulement après l'analyse du contenu des entretiens et non au départ, à partir de données descriptives :

> Abouda & Baude *2008:135, « Du Français Fondamental aux ESLO[39]»)*
>
> *A ce tâtonnement (technique et théorique) dans la recherche du recueil de français spontané en interaction, s'est ajouté celui d'une sociolinguistique applicable à l'enquête.*
>
> *L'entretien en face-à-face a été élaboré sur la base de 3 questionnaires : le premier (ouvert) devait permettre de recueillir les positions du témoin sur son expérience personnelle et divers types de discours déterminés. Le second, semi-fermé, intitulé "questionnaire socio-linguistique" et confié à un élève de Pierre Bourdieu – Bernard Vernier –, s'il porte encore les traces des théories de Bernstein sur la langue, enferme un recueil des représentations du témoin sur la norme linguistique et culturelle. La sociologie naissante de Bourdieu se rencontre également dans le troisième questionnaire, fermé, qui porte, parallèlement à l'état civil, sur les pratiques déclarées des habitudes culturelles.*
>
> *Cette importance donnée au capital culturel s'est également concrétisée dans l'élaboration d'une nouvelle grille, l'échelle AM (de son concepteur Alix Mullinaux, qui comprend cinq agrégats (notés de A à E). Complémentaire à celle de l'INSEE, cette grille tente de rendre compte, parallèlement aux critères habituels, des pratiques et références culturelles ainsi que de la mobilité géographique potentielle des témoins.*
>
> *Ce n'est pas le moindre des intérêts du projet ESLO que de porter les traces de la fin d'une sociolinguistique militante (et naïve) qui se bornait à corréler variations sociales et hiérarchisation de la compétence linguistique, et les prémices d'une nouvelle sociologie qui donnera toute sa place à la distinction apportée par le capital culturel.*

[38] DESROSIERES, A. (2008). *L'argument statistique*.
Eshkol-Taravella et al., "Un grand corpus oral « disponible »."
[39] ABOUDA, L., & BAUDE, O. (2005). « Du Français Fondamental aux ESLO ».

La difficulté d'une classification des locuteurs est réelle et il n'est pas anodin de noter que dès le début de l'entreprise ESLO, la sociologue avait rencontré des difficultés dues au caractère franco-britannique de l'équipe. Elle a ainsi effectué un très gros travail d'alignement des catégories entre la nomenclature française et anglaise. Derrière ce problème s'inscrivent les prémisses d'une véritable réflexion sur les effets de mutualisation des données de la recherche au niveau européen. Il s'agit donc d'une réflexion en avance sur son temps si on se réfère à des études actuelles sur une nomenclature européenne des catégories socioprofessionnelles (Filhon et al. 2013[40]). J'ai d'ailleurs été confrontés à un problème similaire lorsque, pour faciliter l'exploitation du travail d'Alix Mullineaux, j'ai réalisé en collaboration avec trois auteures la traduction en français de la version initiale de l'article. Cette traduction a été très délicate à réaliser justement sur les problèmes d'alignement de la terminologie des métiers entre la version anglaise de l'époque et une version française contemporaine. Ces difficultés attirent l'attention sur les méthodes de comparaison de corpus issues de structures sociales différentes soient géopolitiquement soit chronologiquement.

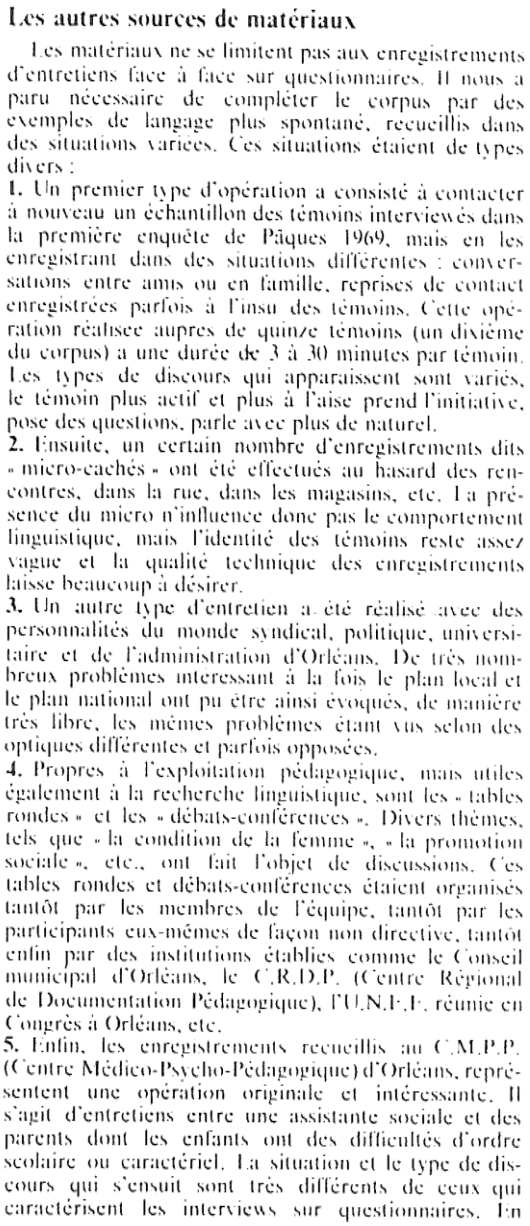
Il reste que l'initiative d'Alix Mullineaux reste inachevée sur un point fondamental : la prise en charge d'une classification qui dépasse les critères classiques de la socio-économie pour intégrer le capital culturel et les comportements sociaux dont on peut penser que le langage, en tant que fait social, y aurait toute sa place. C'est d'ailleurs une des forces à relever du projet ESLO, son caractère novateur et ses échecs n'empêchent pas un usage scientifique pertinent pour peu que les conditions de production du corpus puissent être discutées et interrogées.

3.1.4 ESLO et la variation diaphasique corpus_architecture] [retour]

La variation diaphasique est abordée d'une manière beaucoup moins rigoureuse et systématique dans ESLO. On peut déceler deux approches parallèles ou parfois sécantes. D'une part le corpus est constitué de différents modules consacrés à différentes situations de communication, et d'autre part certains locuteurs du panel ont été enregistrés systématiquement dans ces différentes situations. Il s'agit d'une tentative d'ouverture vers la linguistique interactionnelle et la variation diaphasique :

Blanc & Biggs, *Le français dans le monde* page 19.

[40] FILHON, A. et al. (2013). « Un projet de nomenclature socioprofessionnelle européenne ».



Les autres sources de matériaux

Les matériaux ne se limitent pas aux enregistrements d'entretiens face à face sur questionnaires. Il nous a paru nécessaire de compléter le corpus par des exemples de langage plus spontané, recueillis dans des situations variées. Ces situations étaient de types divers :

1. Un premier type d'opération a consisté à contacter à nouveau un échantillon des témoins interviewés dans la première enquête de Pâques 1969, mais en les enregistrant dans des situations différentes : conversations entre amis ou en famille, reprises de contact enregistrées parfois à l'insu des témoins. Cette opération réalisée auprès de quinze témoins (un dixième du corpus) a une durée de 3 à 30 minutes par témoin. Les types de discours qui apparaissent sont variés, le témoin plus actif et plus à l'aise prend l'initiative, pose des questions, parle avec plus de naturel.

2. Ensuite, un certain nombre d'enregistrements dits « micro-cachés » ont été effectués au hasard des rencontres, dans la rue, dans les magasins, etc. La présence du micro n'influence donc pas le comportement linguistique, mais l'identité des témoins reste assez vague et la qualité technique des enregistrements laisse beaucoup à désirer.

3. Un autre type d'entretien a été réalisé avec des personnalités du monde syndical, politique, universitaire et de l'administration d'Orléans. De très nombreux problèmes intéressant à la fois le plan local et le plan national ont pu être ainsi évoqués, de manière très libre, les mêmes problèmes étant vus selon des optiques différentes et parfois opposées.

4. Propres à l'exploitation pédagogique, mais utiles également à la recherche linguistique, sont les « tables rondes » et les « débats-conférences ». Divers thèmes, tels que « la condition de la femme », « la promotion sociale », etc., ont fait l'objet de discussions. Ces tables rondes et débats-conférences étaient organisés tantôt par les membres de l'équipe, tantôt par les participants eux-mêmes de façon non directive, tantôt enfin par des institutions établies comme le Conseil municipal d'Orléans, le C.R.D.P. (Centre Régional de Documentation Pédagogique), l'U.N.F.F. réunie en Congrès à Orléans, etc.

5. Enfin, les enregistrements recueillis au C.M.P.P. (Centre Médico-Psycho-Pédagogique) d'Orléans, représentent une opération originale et intéressante. Il s'agit d'entretiens entre une assistante sociale et des parents dont les enfants ont des difficultés d'ordre scolaire ou caractériel. La situation et le type de discours qui s'ensuit sont très différents de ceux qui caractérisent les interviews sur questionnaires. En

Le premier point est resté embryonnaire et le projet d'enregistrer les mêmes personnes dans différents contextes se réduit souvent aux enregistrements pré et post entretien. Seul le premier des témoins (BA725) sera enregistré neuf fois :

- en entretien
  - 001 : http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-FRA_ESLO1_1_SOUND ou ark:/87895/1.17-355009 ou hdl:10670/1.ed5v31,
- en micro caché lors de discussion pré entretien :
  - 201 : http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO1_ENTCONT_201 ou ark:/87895/1.17-483233,

        - 202 : http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO1_ENTCONT_202 ou ark:/87895/1.17-483235
        - 203 : http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO1_ENTCONT_203 ou ark:/87895/1.17-483236
    - lors d'un diner familial
        - 270 : Enregistrement en cours de transcription (acoustique très mauvaise)
    - au téléphone (micro caché)
        - 301 : Enregistrement non transcrit
        - 302 : http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO1_TEL_302 ou ark:/87895/1.17-483656
    - sur son lieu de travail (visite des abattoirs)
        - 601 : Enregistrement en cours de transcription partielle (seuls quelques passages sont audibles)
        - 602 : Enregistrement en cours de transcription partielle (seuls quelques passages sont audibles)

D'autres témoins seront enregistrés dans quelques situations différentes (micro caché, réunions) mais on peut supposer que, devant l'ampleur de la tâche, les chercheurs n'ont pas réussi à aller au-delà et ils se sont contentés de quelques opportunités.

Ainsi la forme réelle du corpus ESLO est la suivante :

Enregistrements libres : 36 enregistrements des témoins dans des situations sociales ou professionnelles faits en l'absence des chercheurs (but : comparaison avec des contextes non structurés ou structurés par les témoins)

- "Reprises de contact" : 43 reprises informelles avec les témoins enregistrés à leur insu (but : comparaison de témoignages pris en situation d'interview et dans un contexte moins structuré).
- Enregistrements divers + témoins inconnus : 84 visites d'atelier, marché, magasins, etc.(but : réunir des exemples d'une parole publique différente).
- Communications téléphoniques : 51 communications (but : situation de communication particulière, corpus spécifique de cette forme orale médiatisée).
- Conférences-débats ou discussions : 29 (but fournir des exemples de parole publique, qu'il s'agisse des conférences et des interventions qui les suivent, ou de la trame de la discussion à plusieurs participants).

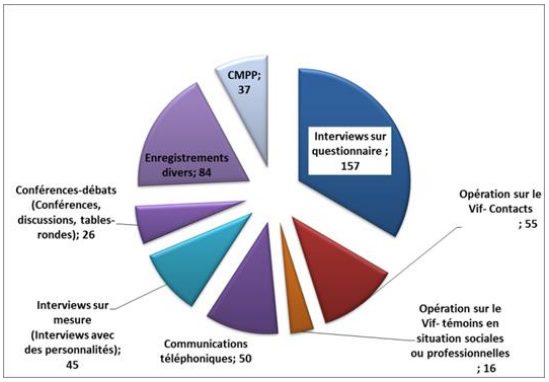
- Interviews au Centre Médico-Psychopédagogique : 41 interviews de parents d'élèves et assistantes sociales. (but : inversion de la motivation interviewer / interviewé).
- Interviews "sur mesure" : 46 interviews avec des personnes choisies selon leur rôle dans la "microsociété" orléanaise (but la constitution d'un portrait sonore de la ville ; sur le plan linguistique des témoignages de personnalités publiques parlant de leur rôle et de leurs activités.

http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000095934
http://cocoon.huma-num.fr/exist/crdo/meta/crdo-COLLECTION_ESLO1

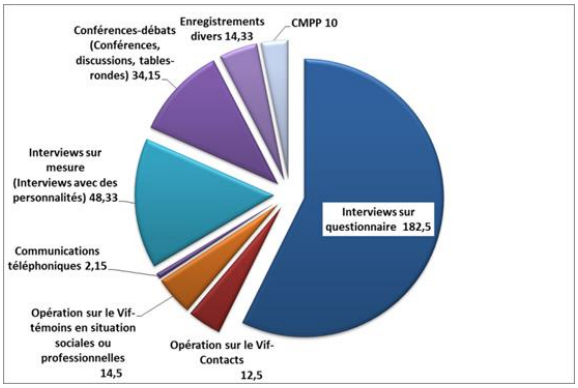
Ces différents modules posent un problème de classification induit par la gestion du corpus.

Variations et effet de gestion sur la variabilité du corpus

Voici une représentation de l'architecture du corpus en nombre de documents sur la base de la typologie utilisée par les auteurs :



Et la représentation de l'architecture en durée d'enregistrement :



Nous pouvons constater, par la lecture rapide de ce simple graphique, qu'il y a une forte différence entre la présentation d'un corpus équilibré en genre, et la réalisation du corpus où les interviews sur questionnaire représentent la grande majorité du corpus. Or nous ne disposons pas d'informations sur le degré de réalisation obtenue par les auteurs du corpus. Est-ce que cette architecture réalisée est conforme à leur souhait ? Nul ne le sait mais le corpus existe maintenant sous une forme d'objectifs et sous une forme concrète constituée de « données ».

La phase de reprise de gestion du corpus numérisé apporte ses propres effets de classification :

Catalogue original (Longman 1974)								Site ESLO 2014
Catégorie des modules	**Description**	**But**	**Nbre**	**Durée**	**Catégorie des sous-modules**	Nbre	**Côte**	**Modules**
Interviews sur questionnaire	Interviews face-à-face sur des questionnaires standardisés, avec un échantillon statistique aléatoire choisi d'après la liste INSEE du recensement de la population 1968.	La constitution d'une gamme sociologiquement représentative de témoignages à contenu constant	**157**	**18 2,5**			001-173	**Entretien**
Opération sur le Vif-Contacts	(Prises de contact, reprises de contact, ouverture et clôture des entretiens, enregistrées à l'insu du témoin)	La comparaison des témoignages pris en situation d'interview et dans un contexte moins structuré Comparaison avec des contextes non-structurés ou structurés par les témoins	**55**	**12, 5**	Reprises de contact avec témoins	35	201-239	**Contact**
					Prises de contact pour l'interview	3	241-243	
					Avant enregistrement	10	250-268	**Ouverture de l'entretien**
					Après enregistrement	7		**Clôture de l'entretien**
Opération sur le Vif-témoins en situation sociales ou professionnelles	(Enregistrements de témoins Insee dans des situations sociales ou professionnelles, faits en l'absence des chercheurs)		**16**	**14, 5**	Repas	10	270-281	**Repas**
					Réunions	4	294-297	**Réunion**
					Magasins	1	290	**Magasin**
					Divers	1	292	**Divers**
Communications téléphoniques		Fournir des exemples de témoins INSEE dans une nouvelle situation de communication, et un corpus téléphonique supplémentaire servant éventuellement à l'étude de cette forme de dialogue médié	**50**	**2,1 5**			301-373	**Appel téléphonique**
Interviews sur mesure (Interviews avec des personnalités)	Interviews "sur mesure" avec des individus choisis selon leur rôle dans la "microsociété" orléanaise	La constitution d'un "portrait" sonore de la ville ; sur le plan linguistique, des témoignages de personnalités publiques parlant de leur rôle et de leurs activités	**45**	**48, 33**	Education	7	401-409	**Interview de personnalités**
					Industrie	17	414-431	
					Syndicats/politique	7	435-441	
					Société/Culture	10	445-454	

					Divers	4	460-463	
Conférences-débats (Conférences, discussions, tables-rondes)	Conférences-débats ou discussions à plusieurs participants (les dernières comportant souvent des témoins INSEE)	Fournir des exemples de parole publique, qu'il s'agisse de conférences et des interventions qui les suivent, ou de la trame de la discussion à plusieurs participants	**26**	**34, 15**	Manifestation	1	501	**Conférences**
					Enseignements-conférences	4	502-505	
					Enseignement : discussion de groupes divers	13	511-523	**Réunion**
					Syndicats : réunions	3	541-543	
					Tables-rondes : divers	4	551-554	
					Conversation avec famille	1	555	**Repas**
Enregistrements divers	Enregistrements divers comportant des témoins inconnus (visite d'atelier, marché, magasins, etc.)	Réunir des exemples d'une parole publique différente : celle du marchand ou du camelot, par exemple	**84**	**14, 33**	Visites	7	601-607	**Visites**
					Enregistrements piégés dans la rue	10	609-618	**Marché**
					Enregistrements piégés dans la rue	1	619	**Divers**
					Tentatives d'achats d'articles plus ou moins exotiques	15	621-635	**Magasin**
					Achats	28	640-667	
					Renseignements divers	5	670-675	
					Renseignements divers	1	676	**Divers**
					Porte à porte à la recherche	14	680-693	
					Divers	3	695-699	
CMPP	Interviews au Centre Médico-Psychopédagogique (parents d'élèves et assistante sociale)	Réunir un corpus servant de contrôle et de contrepoids aux interviews du groupe 1. Alors que les chercheurs de l'ESLO sont venus chez les témoins INSEE solliciter des entretiens, au CMPP, ce sont des interviewés eux-mêmes qui sont venus se présenter. Dans le premier cas, la dynamique de l'interview venait du chercheur ; dans le deuxième, elle venait plutôt des parents qui venaient consulter l'assistante sociale. Pour résumer la situation, on peut dire que dans chaque ensemble d'enregistrements, le contenu reste plus ou moins constant, mais dans le rapport intervieweur/interviewé la motivation se trouve inversée.	**37**	**10**			701-752	**Consultation CMPP**
Total			**470**	**31 8h 30**				

Ce tableau expose les effets de classification des modules du corpus dans le cadre des opérations de gestion de celui-ci. Entre les objectifs annoncés, ceux qui sont implicites, les informations du catalogue, les codes utilisés et les modifications pour répondre à des objectifs de comparaison, on constate la forte variabilité du corpus selon des points de vue aussi simples.

Ainsi l'enregistrement 555 qui est un repas organisé pour répondre à un objectif de discussion collective est catégorisé comme « repas » dans les outils actuels de gestion du site alors qu'il était catégorisé « conférences, réunions » pour les auteurs. Cette « recatégorisation » provient d'une modification du champ linguistique au sein duquel les données dites « non provoquées par le chercheur » prennent une place bien plus légitime en 2015 qu'en 1968. Ainsi un module ESLO2 est dédié aux repas et une thèse est en cours sur ce corpus au sein du laboratoire. Ce module est d'ailleurs l'un des plus recherchés actuellement par les linguistes qui se rendent sur le site de diffusion du corpus. Enfin ce module ouvre un dialogue avec la base de données CLAPI (Corpus de Langues Parlées en Interaction).

Il convient donc de ne pas ignorer les effets de la simple gestion des données sur la forme même que revêt un corpus. Si la linguistique variationniste entend la nature sociale de la langue dans ses variations internes elle doit ne pas être sourde aux contextes de production et de traitement des données scientifiques.

3.1.5 Un portrait sonore disponible ? [Bonnes pratiques] [retour]

Les 471 (487 selon les effets de catégorisation sur le comptage) documents sonores représentant environ 315 heures d'enregistrement aboutissant à une estimation par les auteurs de 4,5 millions de mots, sont décrits dans un catalogue contenant des informations précises sur chaque enregistrement ainsi qu'une indexation riche. Enfin, une partie du corpus a été transcrite de manière manuscrite puis tapuscrite par l'équipe. Ces documents s'ajoutent aux différentes fiches et résumés qui accompagnent le corpus.

L'ensemble du corpus pouvait être repris par d'autres chercheurs. Cette disponibilité du corpus était indiquée dès les premières pages du catalogue comme on l'a déjà mentionné :

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

Nous le verrons, le bilan de cette disponibilité est mitigé, mais il faut reconnaitre au projet ESLO le mérite des efforts faits pour le rendre accessible à tout chercheur. Notons simplement qu'en 1974, les choix de l'équipe étaient réduits à l'indication dans le catalogue que les bandes originales pouvaient être copiées. Deux éléments sont marquants : la conservation et l'accès par les bibliothèques n'a visiblement pas été envisagé et l'informatisation du corpus (Cf. Le corpus de Montréal à la même époque) ne semblait pas constituer une alternative à la diffusion mais seulement un outil de traitement des données. De fait le monde numérique de l'époque n'est pas encore, chez les chercheurs en

linguistique, celui des réseaux ni même celui du traitement automatique du langage, mais celui des langages de machine.

3.1.6 Un contexte scientifique peu disponible [archivage] [retour]

L'expérience de la reprise d'ESLO quarante ans après montre la difficulté de disposer du contexte scientifique d'un projet de corpus. De nombreux éléments méthodologiques et théoriques sont absents ou implicites.

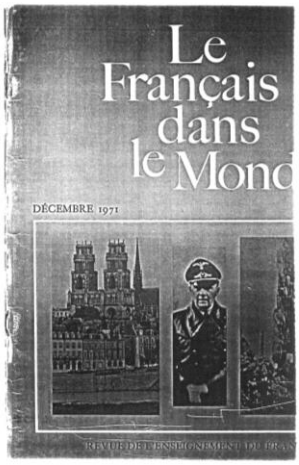
ESLO a donné lieu à très peu de documents papiers contextuels (Cf. numérisation du corpus » et à aucune archive scientifique exceptées les publications suivantes :

- Le catalogue des enregistrements. Document de 265 pages réalisé par Jonna Lonergan et publié en 1974 (11280/deecd9b4)[41]avec la présentation suivante :

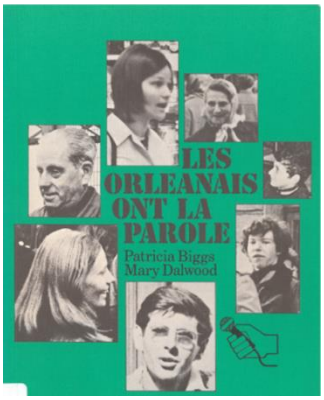
> « *Ce catalogue a été produit, dans le cadre de l'ESLO, par Jonna Lonergan, BA, MA, avec la collaboration de Jack Kay et John Ross, qui prennent la responsabilité des lacunes éventuelles.*
>
> *Mise en page et composition IBM de Joëlle Adams.*
>
> *Couverture et pages de titre par Elma Harvey.*
>
> *Reproduction photo-offset réalisée par l'imprimerie TECHNIQUE, A.G. Leach & Company Limited, Colchester, U.K.*
> *Orléans Archive.*
>
> Colcherster 1974»

- L'article de présentation « L'enquête socio-linguistique sur le français parlé à Orléans » de Michel Blanc et Patricia Biggs, publié dans la revue : *Le français dans le monde* en 1971 (Blanc&Biggs 1971 : 11280/97b0c99a)[42]

[41] https://www.nakala.fr/data/11280/deecd9b4
[42] https://www.nakala.fr/data/11280/97b0c99a



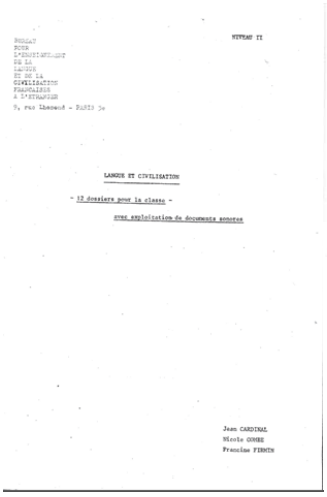
- Les manuels d'enseignement du Français « Les Orléanais ont la parole » publiés par Patricia Biggs et Mary Dalwood (1976).



- L'article d'Alix Mullineaux "The Problems of Classifying the Population Sample in the Socio-Linguistic Survey of Orléans (1969) in Terms of Socio-Economic, Social and Educational Categories" (Mullieaux et Blanc 1982)[43] sur l'échantillonnage du panel des entretiens:

- Les dossiers du BELC "Langue et civilisation, 12 dossiers pour la classe avec exploitation de documents sonores »

[43] MULLINEAUX, A., & BLANC. (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 ».



LANGUE ET CIVILISATION

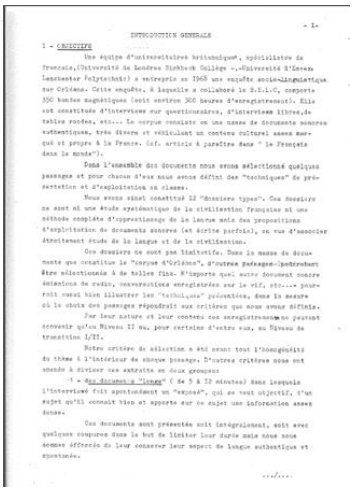
- 12 dossiers pour la classe -

avec exploitation de documents sonores

Jean CARDINAL
Nicole COMBE
Francine FERRUS

TABLE DES MATIERES

Pages



INTRODUCTION GENERALE

I - OBJECTIFS

Une équipe d'universitaires britannique, spécialistes de français,(Université de Londres Birkbeck College -,Université d'Essex Lancaster Polytechnic) a entrepris en 1968 une enquête socio-linguistique sur Orléans. Cette enquête, à laquelle a collaboré le B.E.L.C, comporte 300 bandes magnétiques (soit environ 300 heures d'enregistrement). Elle est constituée d'interviews sur questionnaires, d'interviews libres,de tables rondes, etc... Le corpus constate en une masse de documents sonores authentiques, très divers et véhiculant un contenu culturel assez marqué et propre à la France. (cf. article à paraître dans " le Français dans le monde").

Dans l'ensemble des documents nous avons sélectionné quelques passages et pour chacun d'eux nous avons défini des "techniques" de présentation et d'exploitation en classe.

Nous avons ainsi constitué 12 "dossiers types". Ces dossiers ne sont ni une étude systématique de la civilisation française ni une méthode complète d'apprentissage de la langue mais des propositions d'exploitation de documents sonores (et écrits parfois), en vue d'associer étroitement étude de la langue et de la civilisation.

Ces dossiers ne sont pas limitatifs. Dans la masse de documents que constitue le "corpus d'Orléans", d'autres passages pourraient être sélectionnés à de telles fins. N'importe quel autre document sonore (émissions de radio, conversations enregistrées sur le vif, etc...) pourrait aussi bien illustrer les "techniques" présentées, dans la mesure où le choix des passages répondrait aux critères que nous avons définis.

Par leur nature et leur contenu ces enregistrements ne peuvent souvent qu'un Niveau II ou, pour certains d'entre eux, au Niveau de transition I/II.

Notre critère de sélection a été avant tout l'homogénéité du thème à l'intérieur de chaque passage. D'autres critères nous ont amené à diviser ces extraits en deux groupes:

1 - des documents "longs" (de 5 à 12 minutes) dans lesquels l'interviewé fait spontanément un "exposé", qui se veut objectif, d'un sujet qu'il connait bien et apporte sur ce sujet une information assez dense.

Ces documents sont présentés soit intégralement, soit avec quelques coupures dans le but de limiter leur durée mais nous nous sommes efforcés de leur conserver leur aspect de langue authentique et spontanée.

.../....

3.2 D'ESLO à ESLO1 « numérique » : la transformation d'un objet scientifique [Numérisation] [retour]

3.2.1 Maniabilité du corpus et rayonnement d'ESLO

Le corpus d'Orléans (ESLO) n'a connu ni une diffusion massive, ni un impact fort dans les recherches en linguistique. Citons néanmoins la thèse de De Jong sur la «sociophonologie de la liaison[44]» et le livre de Greidanus sur les constructions verbales en français parlé[45].

Comment expliquer qu'un corpus, unique en son genre et répondant à une actualité scientifique grandissante, la description du français parlé, ait été aussi peu utilisé ? Différentes causes expliquent ce fait, mais, parmi elles, la difficulté à «manier» un corpus de cette taille est un frein considérable. En 1974, date de la parution du catalogue, le corpus d'Orléans se présente sous la forme de plus de 300 bandes magnétiques qu'il faut lire à l'aide d'un magnétophone à bandes.

[44] De Jong, D. (1994). « La sociophonologie de la liaison orléanaise ».
[45] Greidanus, T. (1990). *Les constructions verbales en français parlé: étude quantitative et descriptive de la syntaxe des 250 verbes les plus fréquents*.

Le catalogue papier est le seul outil qui permet de fouiller le corpus, l'autre méthode reposant sur l'écoute du corpus (un chercheur qui l'écouterait huit heures par jour y passerait plus d'un mois). Les transcriptions réalisées à l'époque étaient très partielles, d'une faible qualité et surtout uniquement disponibles en version manuscrite ou imprimée. Si ESLO était précurseur sur de nombreux aspects, les auteurs n'ont pas du tout repéré la naissance de l'informatique. A la même époque, le corpus de Montréal faisait un choix rigoureusement opposé (Thibault & Vincent 1990)[46].

Le travail fait par Piet Mertens dans le cadre des projets ELILAP-ELICOP[47], qui a permis la mise en ligne d'une grande partie du corpus, transcrit et étiqueté, a été une avancée considérable qui a permis un large développement des travaux. Toutefois, pour des raisons purement techniques, les enregistrements n'étaient pas disponibles et ce grand corpus oral n'existait toujours pas sous sa forme sonore.

3.2.2 Numérisation d'ESLO [retour]

L'histoire du fonds ESLO est symptomatique de la gestion des archives sonores par la linguistique française. Celui-ci n'a pas été déposé dans une institution de conservation (le service des Archives sonores à la BnF existe pourtant depuis le programme inaugural de Ferdinand Brunot en 1911), et il avait d'ailleurs disparu de France jusqu'à la fin des années 1980. A cette époque le département de français de l'université anglaise qui avait initié le projet ferma et les responsables contactèrent l'université d'Orléans afin de proposer de transmettre les archives dont les bandes magnétiques, faute de quoi l'ensemble serait détruit. Une mobilisation du professeur de linguistique Gabriel Bergounioux, du professeur d'anglais, Jean Baraduc, et du maire d'Orléans qui se trouvait être également enseignant-chercheur en linguistique permit de réquisitionner un véhicule de la ville afin de se rendre, in extremis en Angleterre et de ramener un lot de 7 cartons contenant le précieux corpus avant que celui-ci ne soit mis à la benne. L'université d'Orléans a réalisé une copie sur microcassettes et les documents sont transmis aux Archive départementales.

[46] THIBAULT, P., & VINCENT, D. (1990). *Un corpus de français parlé. Montréal 84 : Historique, méthodes et perspectives de recherche*.
[47] « International Association of Sound and Audiovisual Archives ». (s. d.).

En 2004, le Centre Orléanais de Recherche en Anthropologie et Linguistique (CORAL), laboratoire de l'Université d'Orléans qui deviendra ensuite l'unité mixte de recherche Laboratoire Ligérien de Linguistique (LLL), entreprit la numérisation du corpus avec le soutien du Ministère de la Culture. Nul ne réalise véritablement, à l'époque, que le corpus va entrer dans une nouvelle ère.

Numérisation des bandes *[retour]*

La première opération consista à numériser les bandes magnétiques dont on pouvait craindre une détérioration due au cycle de vie des supports magnétiques sur bandes.

> (Baude & al., *Corpus oraux, guide des bonnes pratiques 2006*, pp148-149) [48]
>
> *ORGANISATION DE LA CHAINE DE NUMERISATION*
> *La numérisation se décompose en plusieurs étapes mais a pour règle de base la meilleure relecture possible du document d'origine.*
> *PREPARATION DOCUMENTAIRE ET PHYSIQUE DES ELEMENTS*
> *(…) CHAINE DE TRANSFERT*
> *(…) NUMERISATION, COMPRESSION*
> *(…) CONTROLE QUALITE*
> *(…) CORRECTION DU SIGNAL*
> *(…) VERSEMENT DANS L'ARCHIVE NUMERIQUE*
> *Ces technologies sont soumises à un cycle d'obsolescence rapide, qui contraindra à des migrations de masse tous les cinq ou six ans environ. D'où la nécessité impérieuse de disposer d'une visibilité financière à moyen terme, au-delà de l'opération de passage au numérique, pour garantir la pérennité des investissements engagés et – surtout – l'accès aux fonds qui ne seront bientôt plus accessibles du tout sous leur forme analogique d'origine.*

En attendant une numérisation de qualité à des fins de pérennisation, une première opération de numérisation a été effectuée afin de faciliter le maniement des enregistrements et de réaliser une transcription alignée sur le signal. Après l'acquisition d'un banc de numérisation et une formation assurée par le responsable du programme des archives sonores du LACITO, une première numérisation dédiée à la transcription fut réalisée par des étudiants vacataires selon une procédure établie par l'équipe. Le format de numérisation fut celui recommandé à l'époque par l'IASA : copie droite des bandes en WAV, stéréo, 44100 Hz, 16 bits.

[48] BAUDE, O. et al. (2006). *Corpus oraux : guide des bonnes pratiques 2006*.

Les bandes numérisés furent conservées par l'intermédiaire d'un quadruple jeu de cédés. Par la suite ces fichiers seront transférés sur des disques durs avant de bénéficier du programme du CNRS sur l'archivage scientifique à partir du projet pilote sur les données orales.

Enfin en 2013, le fond sera déposé à la BnF qui assurera une nouvelle numérisation dans le respect de ses pratiques.

Catalogage *[métadonnées]*

Parallèlement le catalogue papier sera numérisé en format image mais surtout repris intégralement sous la forme d'une base de données mysql. Cette opération permet de conserver et rendre accessible la description des bandes selon les choix originels. Les mêmes informations structurées en base de données offrent des possibilités d'accès aux données totalement différentes. Il est ainsi possible d'explorer le corpus en sélectionnant des champs de description et en les croisant.

Par la suite une seconde phase de catalogage, qui s'appuiera sur les pratiques et normes d'archivage des documents sonores, ouvrira l'accès de cette description du corpus à l'ensemble des catalogues et bases de données documentaires suivant les principes de cette interopérabilité maximale. Selon les pratiques en vigueur cette description se fera à l'aide de différents formats et normes de métadonnées.

> (Baude & al., *Corpus oraux, guide des bonnes pratiques 2006*, pp148-156) [49]
>
> Un document numérique, quel qu'il soit, n'est pas pérennisable sans un minimum de métadonnées associées. Les métadonnées minimales sont celles qui permettront l'identification du contenu, la description complète de son mode de production (description de la chaine de numérisation) et les caractéristiques techniques du format qui permettront d'engager des actions de pérennisation (par exemple la migration vers un autre format) en cas de risque. Pour être exploitables informatiquement, ces métadonnées doivent obéir

[49] Ibid.

> strictement à une formalisation (par exemple dans le langage de balise XML). Afin de limiter au maximum les saisies manuelles, perte de temps pour les techniciens, les métadonnées devront être générées automatiquement en exploitant les informations déjà connues préalablement (celles issues notamment du travail de préparation documentaire).
>
> D'autres métadonnées pourront par ailleurs être ajoutées à loisir selon l'usage : vignettes périodiquement extraites du document comme aide à la consultation ; image numérisée de jaquettes ou de fiches papier associées au document vidéo ; indexation temporelle du contenu ; ou encore (dans le futur) reconnaissance de la voix permettant une recherche « plein texte », etc.
>
> (…) Les métadonnées servent à décrire des ressources (enregistrements, annotations). Ces descriptions peuvent contenir des informations sur la nature physique des ressources (durée de l'enregistrement, format de fichier, etc.), sur les droits associés, sur la situation d'enquête (lieu, date, participants, etc.). Ces métadonnées correspondent aux renseignements que l'on pourrait trouver dans une notice bibliographique de bibliothèque. Il existe un certain nombre de renseignements communs avec ce type de notice, mais les caractéristique propres des corpus oraux, ainsi que les préoccupations particulières des personnes qui les étudient, ont conduit à la définition de champs tels que l'âge du locuteur ou les conditions d'enregistrement, que l'on aura plus de mal à faire entrer dans une notice classique de bibliothèque. Les métadonnées servent principalement à deux choses : à cataloguer et à échanger. Pour que les échanges soient possibles, il convient de normaliser à la fois la forme des métadonnées mais aussi la procédure d'échange.

En effet, plusieurs codages ont été proposés et sont utilisés pour la description des enregistrements et de leurs annotations. La TEI propose d'écrire toutes ces informations dans un en-tête assez détaillé. Pour les ressources du web, Dublin-Core, normalisé ISO-15836 en 2003, propose un jeu de quinze étiquettes qui sont notamment utilisées dans les entêtes des fichiers HTML. Il existe bien sûr les codages pratiqués par les bibliothèques tels que les standards Marc, US-Marc etc. et les instruments de recherche en EAD (Encoding Archive Description) qui se sont adaptés pour coder les nouveaux supports informatiques. Il existe aussi des communautés qui ont proposé des recommandations, comme par exemple OLAC (basé sur du Dublin-Core enrichi et spécifié pour l'adapter aux ressources linguistiques), ou IMDI pour l'infrastructure européenne CLARIN.

Pour le corpus ESLO la numérisation fut l'occasion d'une description selon ces formats dont l'incidence est forte en termes de construction d'un objet de connaissances partagé :

> Abouda & Baude 2008 « Du Français Fondamental aux ESLO »
>
> Dans le cas d'un corpus numérique, il est aisé d'établir des relations entre les données primaires et les principes d'élaboration du corpus, la normalisation et les formalismes choisis, les techniques utilisées et de nombreuses autres informations (ou méta-informations). Or, cette documentation est notamment le lieu pour fournir de précieux renseignements sur la situation de collecte et le profil des témoins. Cette opération a été repérée comme étant fondamentale depuis le développement de la linguistique de corpus : « *La documentation doit couvrir deux volets distincts : les sources utilisées et la responsabilité éditoriale de constitution du corpus d'une part, les conventions d'annotation d'autre part* » (Habert et al. 1997, p.156). Récemment, le langage XML apporte une solution convaincante en séparant les données et les informations sur la structure des données, alors décrites dans l'en-tête du document (recommandations de la TEI).
>
> La gestion de ces informations souvent répertoriées sous le terme de métadonnées rend nécessaire une uniformisation du traitement comme le proposent différentes initiatives centrées sur la gestion, la diffusion et la réutilisation des corpus (EAGLES, OLAC).
>
> Dans le cas du corpus d'ESLO, nous disposons d'un exemple particulièrement intéressant car les métadonnées avaient déjà été répertoriées pour la publication du catalogue en 1974. Or, la transformation du catalogue en une base de données offre des perspectives infinies de requêtes dans d'excellentes conditions. Ainsi, les données sociologiques ont été intégrées à des bases de données relationnelles et deviennent facilement disponibles comme champs que l'on peut croiser avec des requêtes sur la transcription et l'annotation des données linguistiques.

A titre anecdotique, mais cela ne manque pas de saveur, nous pouvons mentionner le choix du codage dans le catalogue ESLO1 :

> « *Les systèmes employés pour identifier les témoins et les bandes lors du travail sur le terrain présentaient certains inconvénients : il existait des systèmes différents pour différentes sections du corpus et tous les codes étaient volontairement redondants afin de réduire au minimum le risque d'erreurs (…). Il convenait donc de chercher un système simple, capable de*

> *recouvrir le corpus tout entier et en vue de* ***l'éventuelle mise en ordinateurs des données, peu encombrants afin de libérer le minimum de bits***[50]*. Nous avons donc pris comme point de départ les huit grandes catégories d'enregistrements en les regroupant selon les codes numériques à trois chiffres* ». (Lonergan 1974 :2)

Transcription *[transcription]*

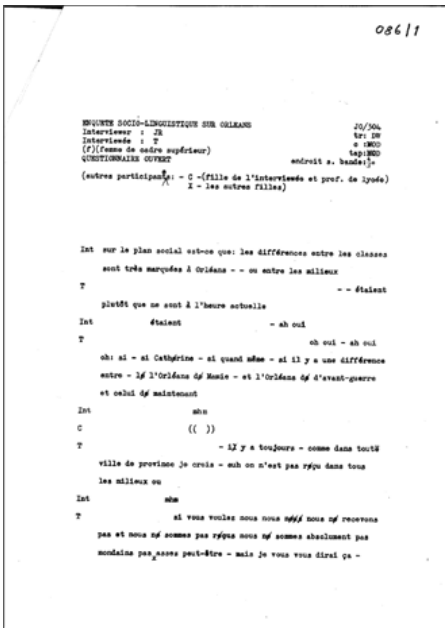
L'opération la plus lourde dans le cadre de la numérisation du corpus est incontestablement la réalisation de la transcription du corpus à l'aide d'outils permettant la synchronisation sur le signal. L'objectif est de disposer d'un outil de fouille du corpus permettant de rechercher un mot ou une chaine de caractères et de se positionner immédiatement sur le segment sonore correspondant afin d'écouter l'enregistrement. Derrière cette étape particulièrement structurante se trouvent différents niveaux d'intervention.

> (Baude & al., *Corpus oraux, guide des bonnes pratiques 2006* :73) [51]
>
> La transcription est une pratique qui, loin de se limiter à un exercice technique de reproduction, intègre de nombreux enjeux théoriques et interprétatifs (déjà Ochs, 1979). Dans le passage de l'oral à l'écrit graphico-visuel, de nombreuses opérations de catégorisation sont effectuées, soit quant aux formes linguistiques, segmentées visuellement en unités (Blanche-Benveniste & Jeanjean, 1987 ; Mondada, 2000), soit quant à l'identité des locuteurs eux-mêmes (Mondada, 2003).

L'ensemble des opérations et analyses est décrite dans un chapitre dédié à cette étape. Nous pouvons toutefois résumer les points principaux du traitement des données sonores numérisées dans le cadre d'un corpus oral.

Précisons néanmoins que les transcriptions réalisées par l'équipe d'ESLO existent sous une forme manuscrite et parfois tapuscrite. Celles-ci n'ont pas été numérisées et la phase de transcription a été reprise à zéro à l'exception des documents qui avaient été transcrits dans le cadre des projets des équipes d'Amsterdam et de Louvain (cf. infra).

[50] Souligné par nous
[51] BAUDE, O. et al. (2006). *Corpus oraux : guide des bonnes pratiques 2006*.



086/1

ENQUETE SOCIO-LINGUISTIQUE SUR ORLEANS
Interviewer : JR
Interviewée : T
(F)(Femme de cadre supérieur)
QUESTIONNAIRE OUVERT

(autres participants : - C -(fille de l'interviewée et prof. de lycée)
X - les autres filles)

Int sur le plan social est-ce que: les différences entre les classes sont très marquées à Orléans - - ou entre les milieux
T - - étaient plutôt que ne sont à l'heure actuelle
Int étaient - ah oui
T oh oui - ah oui oh: si - si Cathérine - si quand même - si il y a une différence entre - le l'Orléans d' Monic - et l'Orléans d' d'avant-guerre et celui d' maintenant
Int mhm
C (( ))
T - il y a toujours - comme dans toute ville de province je crois - euh on n'est pas reçu dans tous les milieux ou
Int même
T si vous voulez nous nous nous nous recevons pas et nous n' sommes pas reçus nous n' sommes absolument pas mondains pas assez peut-être - mais je vous vous dirai ça -

Exemple de transcription tapuscrite ESLO

La transcription numérique entreprise dans le cadre du projet des ESLO transforme l'objet selon les caractéristiques suivantes :

- Synchonisation son-texte

Les enregistrements sonores numérisés permettent une synchronisation du son et du texte à l'aide de jalons temporels assurant la relation entre des segments du fichier son et des segments de la transcription réalisée. Il s'agit d'une avancée technologique majeure qui bouleverse la nature même de l'objet d'étude puisque le son redevient l'objet central d'étude.

- La transcription première des annotations

Ce bouleversement a un effet sur le statut de la transcription. Jusque-là celle-ci devenait la donnée primaire analysée. Avec le recours systématique au son, elle peut être conçue comme une simple annotation et de ce fait doit être catégorisée comme telle dans le process.

> (Eshkol-Taravella et al 2011 :25 « Un grand corpus oral « disponible » : l'expérience du Corpus d'Orléans 1968-2012[52] »)
>
> Nous considérons le processus d'annotation comme porteur d'une interprétation. Ainsi, il n'y a pas qu'une version de corpus annoté mais plusieurs versions – existantes ou potentielles – du même corpus. Les différents annotateurs humains peuvent interpréter et percevoir

[52] ESHKOL-TARAVELLA, I. et al. (2011). « Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012 ».

différemment les données. Même dans le cas de l'annotation automatique, les annotations diffèrent par les conventions, les techniques, etc. ESLO est un corpus de variations : variations entre le français d'hier et d'aujourd'hui, entre les différents locuteurs, entre les différentes situations d'enregistrement, mais aussi entre les différentes annotations. (…)

Pour nous, la transcription doit être considérée comme un premier niveau d'annotation : le son étant enrichi d'une information orthographique. Dans le cadre d'un corpus sociolinguistique, on souhaite également intégrer des descripteurs de la situation d'interaction, notamment les données sociologiques sur les locuteurs et la description de la situation d'enregistrement.

- Navigation dans des données massives

Avec l'utilisation de cette annotation synchronisée par des jalons temporels, il devient plus facile d'utiliser les outils informatiques, et notamment ceux du TAL, pour naviguer rapidement dans un grand corpus de données sonores.

Abouda & Baude 2006 « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. [53]»

Cette première transcription est conçue comme une transcription de base avec un simple statut d'outil de navigation au sein du corpus sonore et de repérage de phénomènes selon une granularité grossière.

Abouda & Baude 2008 « Du Français Fondamental aux ESLO »

Nous avons conçu cette première transcription à un degré le plus proche du zéro, en lui donnant uniquement le statut d'outil de navigation au sein du corpus sonore. L'outil sélectionné a été Transcriber pour sa simplicité d'utilisation, sa robustesse face à des fichiers longs, et sa sortie en un format de fichier XML qui nous a semblé être une garantie d'interopérabilité.

Les conventions de transcriptions ont donc été réduites au minimum. Cependant, même à ce niveau "zéro", de nombreuses questions restent présentes comme la structuration des segments et leur granularité – qu'est-

[53] ABOUDA, L., & BAUDE, O. (2006). « CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO ».

ce qu'un mot ? une phrase ? un tour de parole ? –, le choix des évènements à transcrire, la gestion des chevauchements et des pauses.

- Variations dues à la méthodologie de corpus numériques

Le gain procuré par la numérisation du son et par sa synchronisation avec une annotation primaire apporte à son tour de nouvelles variations induites par la technologie et la méthodologie. En effet, les conventions de transcriptions ne sont plus contraintes seulement par le cadre théorique mais elles le sont aussi par l'impact des choix concernant l'utilisation d'outils ce qui, à une époque où l'interopérabilité est une notion dominante, devient un enjeu crucial. Ceci d'autant plus que la plupart des outils ont été conçus non seulement pour des corpus écrits mais aussi pour, et par, une linguistique de corpus pour laquelle les conditions de production des données font l'objet de bien moins d'interrogations que les opérations de traitement.

Il en résulte une variation qui n'est jamais définie comme telle par les sociolinguistes. Elle ne fait effectivement pas partie du système et n'est pas accessible aux locuteurs. Mais, paradoxalement, elle est fortement présente dans les corpus d'études. Corpus qui ont souvent pour objectif d'être représentatifs, et plus encore, de conduire les analyses à partir des données.

Documentation annexe

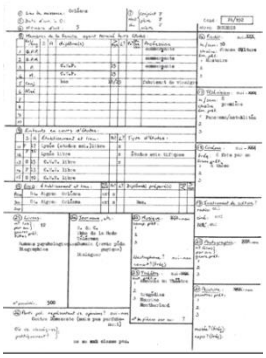
Le fonds ESLO contient également une documentation associée sous la forme d'un dossier par interview sur questionnaire. Il s'agit de différentes fiches manuscrites ou tapuscrites ayant servi à mener les entretiens et à élaborer le catalogue. Elles contiennent des précisions qui apparaissent seulement sur ces fiches, comme par exemple le nom de chaque témoin associé à son identifiant.

- Fiche identité

Cette fiche contient les informations nominatives (noms, adresse) ainsi que les informations fournies par l'INSEE pour la construction de l'échantillon.

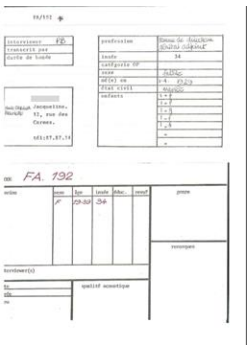
- Fiche sociologique

Fiche contenant les informations sur la famille et son parcours, ainsi que les réponses sur les pratiques culturelles et l'orientation politique déclarée.



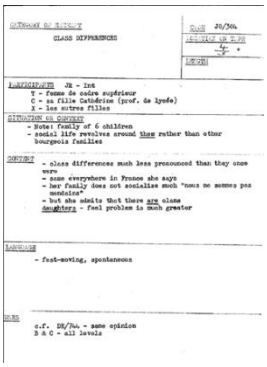
- Fiche entretien

Fiche d'identification de l'entretien sur questionnaire



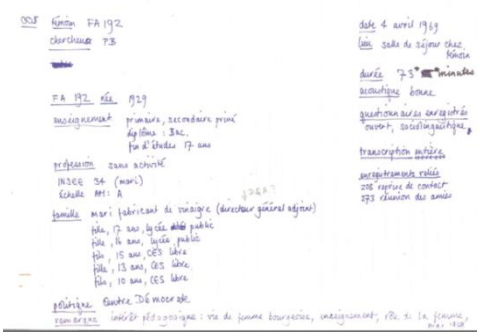
- Fiche méthode de langues

Fiche contenant un repérage des passages pertinents pour une exploitation didactique du corpus.



- Brouillon fiche catalogue

Brouillon manuscrit de la description de l'enregistrement suivant la présentation du catalogue tel qu'il sera édité.

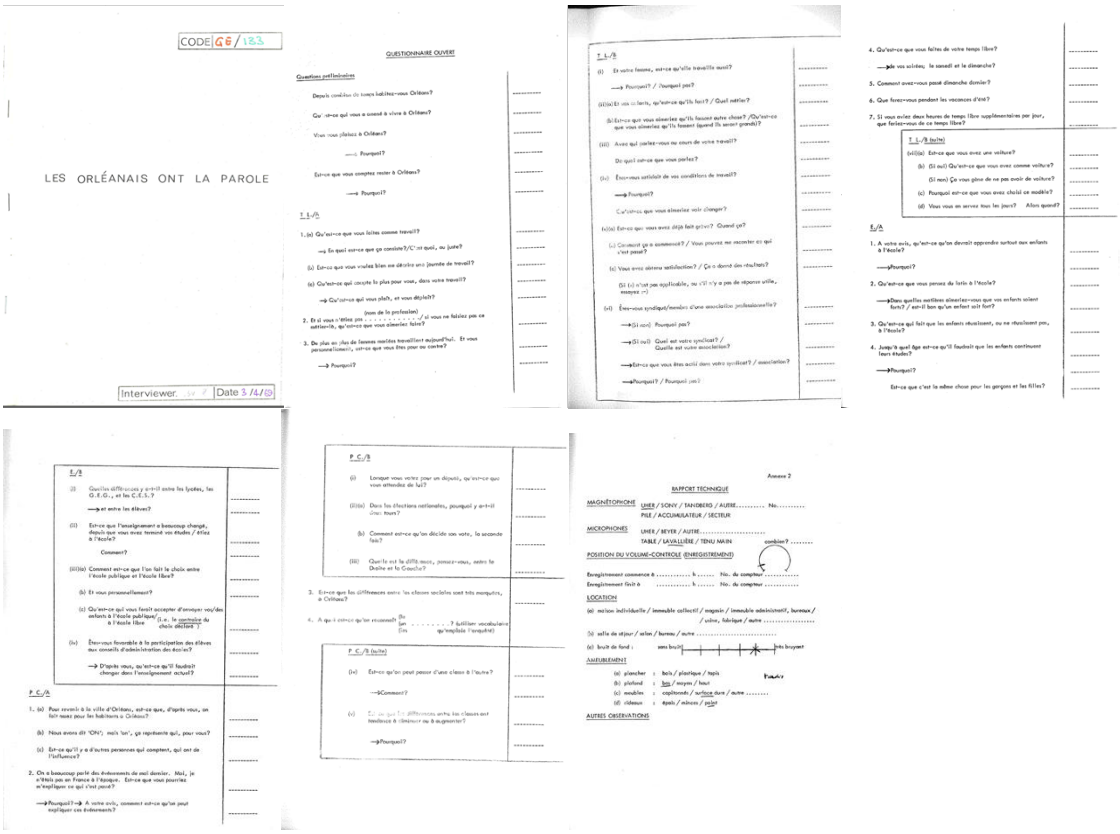


- Fiche interview sur questionnaire

Formulaire utilisé pour la description des entretiens sur questionnaire.

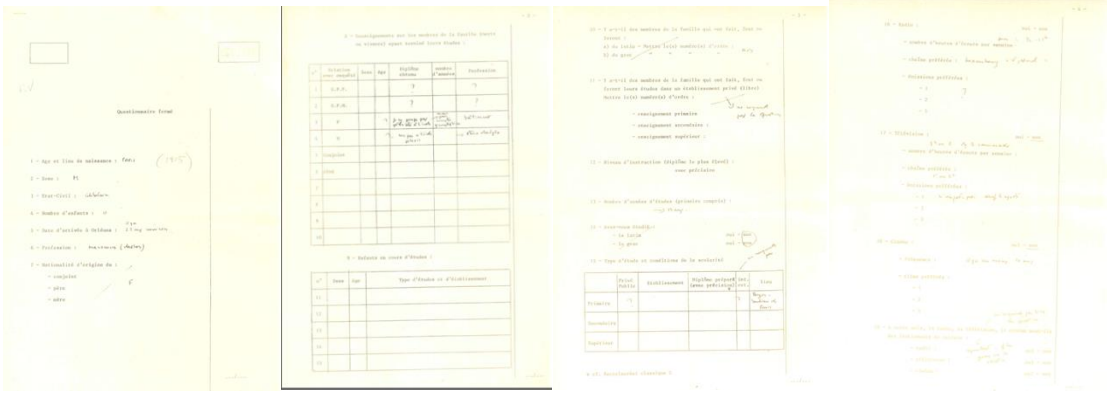
- Fiches questionnaire ouvert

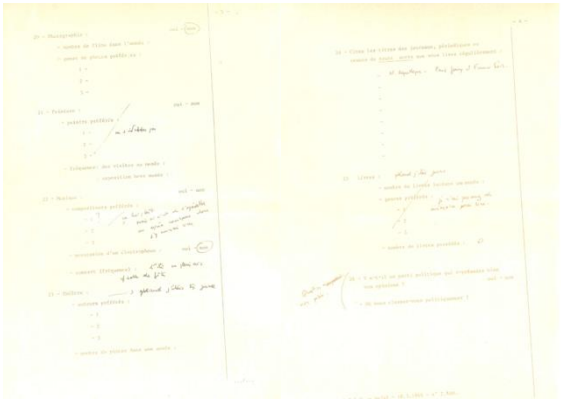
Formulaire en sept pages utilisé pour collecter les réponses aux questionnaires utilisés lors de l'entretien.



- Questionnaire fermé

Fiches en six pages du questionnaire fermé.





- Questionnaire « langue et culture »

Fiche en deux pages de la branche « langue et culture » du questionnaire.

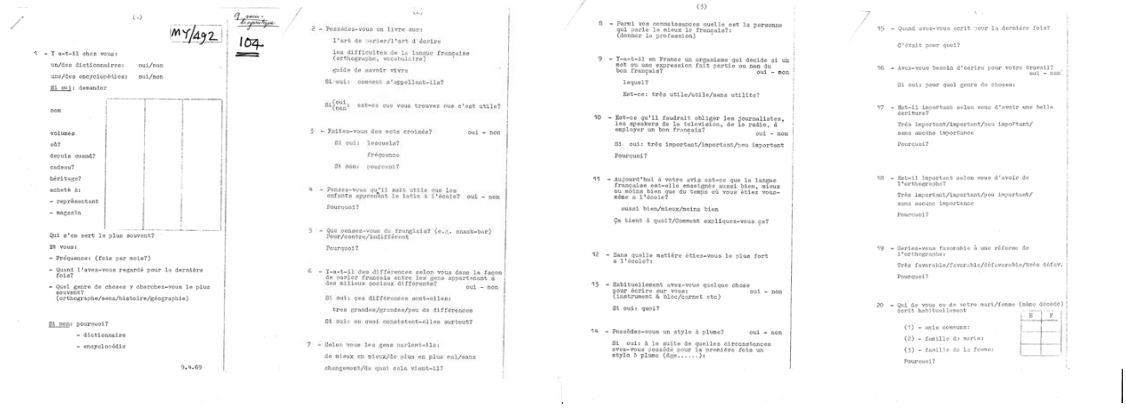
Enquête socio-linguistique sur Orléans

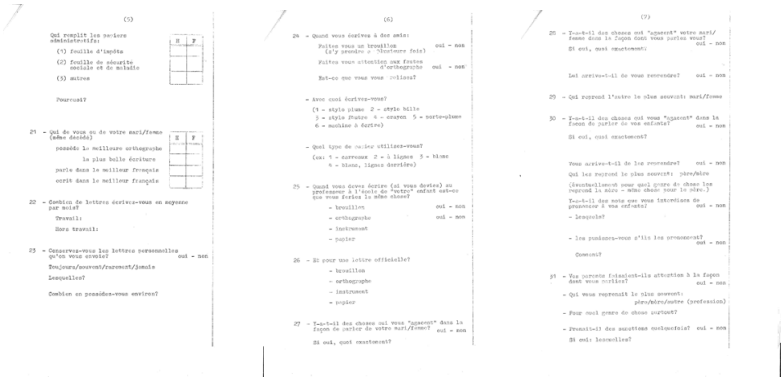
Questionnaire

Branche - langue et culture

- Questionnaire « sociolinguistique »

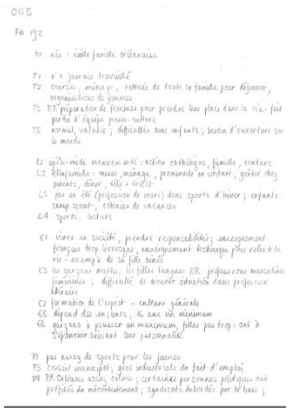
Formulaire en sept pages du questionnaire « sociolinguistique » (titre des auteurs) portant sur les dictionnaires et les pratiques linguistiques (au sens large).





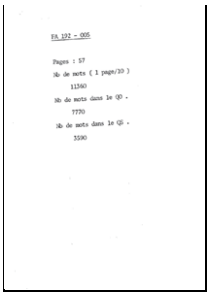
- Indexation de l'entretien

Indexation manuscrite de l'entretien. Brouillon réutilisé dans la fiche du catalogue.



- Fiche « mots »

Fiche présentant un comptage (estimatif ?) du nombre de mots contenus dans un entretien.



Chaque dossier contient donc 12 documents annexes utilisés pour la collecte, la constitution et l'analyse du corpus. La plupart des informations sont reprises dans le catalogue et de ce fait intégrées à la base de données lors de la numérisation de celui-ci. La version papier de ces documents est conservée par la BnF dans le cadre du dépôt du fonds dans son intégralité. Ils sont donc consultables par l'intermédiaire de cette institution.

Il n'existe aucune documentation sur les autres modules du corpus et aucun document contextuel n'a été conservé.

Cet état de fait fige le corpus dans une forme qui est loin d'être satisfaisante pour un projet aussi riche que celui d'ESLO. Le passage au format numérique du corpus fixe un second état qui tamise selon une granularité plus fine encore les données même si la numérisation du fonds en préserve l'intégralité. Il ne faut donc pas confondre numérisation à des fins de conservation et forme numérique.

Enfin cette expérience de l'usage des documents contextuels et des métadonnées d'un corpus nous éclaire sur l'enjeu et les difficultés d'objectivation du corpus en SHS. Nous pouvons ici rejoindre A. Desrosieres (2008) dans cette critique de la relation aux données et aux processus de production de celles-ci :

> *Paradoxe des métadonnées. "D'un point de vue normatif, il semble indispensable de communiquer aux utilisateurs le maximum d'informations détaillées sur les processus de production des données, mais il est vrai aussi que, d'un point de vue descriptif (sans porter de jugement), nombre d'utilisateurs n'ont pas trop envie de trop de métadonnées : l'information 'idéale' est celle qui semble se suffire à elle-même, sans notes en bas de page parasitant le message."* (Desrosieres, 2008 :217[54])

Aspects juridiques *[GBP]*

La numérisation d'ESLO a également déclenché une transformation du corpus à partir de questions juridiques et éthiques. Ce corpus est en effet un cas d'école très particulier et les réponses à ces questions ont largement bénéficié des réflexions du *Guide des bonnes pratiques*[55].

Les principales questions, comme dans la majorité des corpus oraux, se concentrent autour de la gestion des données personnelles et de la propriété intellectuelle.

- Propriété intellectuelle *[GBP]*

La question de la propriété intellectuelle est apriori assez simple à résoudre puisque l'article de référence et le catalogue expriment explicitement une autorisation de réutilisation des données. Nous avons donc choisi d'utiliser des licences creatives commons (CC-BY-SA)[56] qui correspondent à la fois à des pratiques en vigueur et aux objectifs affichés de l'époque.

Dans le cas d'ESLO, elles permettent de respecter la paternité du corpus en assurant une traçabilité de la propriété intellectuelle des opérations postérieures au projet original. Ces décisions ont également une incidence sur les métadonnées.

[54] DESROSIERES, A. (2008). *L'argument statistique*.
[55] BAUDE, O. et al. (2006). *Corpus oraux : guide des bonnes pratiques 2006*.
[56] Site des licences creatives commons

Ainsi l'équipe franco-britannique, et ses deux animateurs (Michel Blanc et Patricia Biggs), sont les auteurs du corpus et une liste nominative des membres intervenants dans chaque document est établie. A l'équipe s'ajoutent les témoins du panel qui, pour des raisons évoquées *infra* ne sont mentionnés que par un code anonyme.

Pour la version numérisée et traitée pour une ré-exploitation et une diffusion dans le cadre du projet des ESLOs, le LLL (ex CORAL) en est le dépositaire et Oliver Baude et Céline Dugua, qui sont les co-responsables du programme, en sont les éditeurs. Les membres de l'équipe intervenant avec des rôles divers (enquêteurs, transcripteurs, etc.) sont également indiqués.

Ces décisions résultent d'une succession de choix réalisés au cours de la chaine de traitement de numérisation. On y décèle la transformation d'un objet simple (un corpus d'enregistrements et de transcriptions) en un objet numérique d'une forme plus complexe comme aboutissement de nombreuses opérations effectuées par des acteurs différents.

- Gestion des données personnelles *[GBP]*

Ce point est beaucoup plus compliqué car l'évolution des pratiques a été importante et les effets technologiques ont considérablement changé en quarante ans. Le fonds ESLO ne contient aucun document spécifique de recueil de consentement des locuteurs. Toutefois, certaines interviews contiennent l'enregistrement de discussions permettant de conclure de manière affirmée que les témoins étaient parfaitement informés de leur participation au programme de recherche et qu'ils ont exprimé systématiquement de manière explicite leur accord pour être enregistrés et pour que cet enregistrement soit utilisé dans le cadre d'une exploitation scientifique et pédagogique. S'ajoute à ceci le fait que les personnes interviewées ont répondu positivement à une sollicitation, ce qui s'apparente à une démarche volontaire.

La gestion de la forme numérisée apporte toutefois un problème d'une tout autre envergure pour deux raisons. D'une part, il était impossible en 1968 d'imaginer une diffusion instantanée, universelle et très facilement accessible par la Toile. D'autre part, des outils de plus en plus élaborés dans le cadre des « big data » offrent une considérable puissance de « fouille » des données. Un consentement de l'époque ne pouvait prendre en compte ces conditions d'utilisation du corpus.

Deux solutions sont toutefois disponibles pour répondre à ces difficultés :

Premièrement l'anonymisation, qui comme le rappelle le *Guide des bonnes pratiques*[57], fait sortir le document du cadre des données personnelles. L'équipe d'ESLO avait pris soin de procéder à un codage des témoins utilisé dans tous les documents du corpus, à l'exception d'une fiche présente dans chaque dossier. Ce codage a été systématiquement repris dans la version numérisée : transcription, BDD, métadonnées. Les adresses, noms et relation avec l'identifiant du codage ont été conservés dans une base de données logistiquement et physiquement séparée comme le recommande le *Guide des bonnes pratiques*. La version numérisée de la fiche d'origine est gérée par la BnF selon ses pratiques de restriction d'accès aux données sensibles des archives publiques.

Deuxièmement, la responsabilité assumée par l'équipe des ESLOS en charge de la diffusion du corpus numérisé, dans le respect d'une démarche éthique, les met à même d'évaluer le consentement des auteurs. L'absence de contenus préjudiciables à l'auteur ou à quiconque est vérifiée et la possibilité de retirer tout extrait ou document entier à la demande d'un auteur ou d'un ayant droit est garantie.

Archivage et accès aux documents numériques *[Archivage]* *[disponibilité français parlé]*

La solution d'archivage et d'accès aux données numériques du corpus ESLO1 a été traitée conjointement avec ESLO2 dans le cadre du programme des ESLOs. Cette solution est détaillée dans un autre chapitre.

Toutefois, ESLO1 a la particularité d'être un fonds sonore enregistré sur des bandes magnétiques dont le cycle d'utilisation est limité. Ce point spécifique a donné lieu à un projet expérimental avec le service des documents sonores du département de l'audiovisuel de la BnF et le CNRS par l'intermédiaire de la plateforme d'archivage des corpus oraux « Cocoon » et de la TGIR Huma-Num. Nous présenterons ici la première partie de cette expérience issue d'un partenariat des plus fructueux[58].

Après avoir été récupéré au service des archives départementales du Loiret où 7 cartons ESLO étaient entreposés, puis numérisés une première fois par le LLL, le fonds « ESLO » a été déposé à la BnF en 2013.

[57] BAUDE, O. et al. (2006). *Corpus oraux : guide des bonnes pratiques 2006*.
[58] Ce travail a été fait par Audrey Viault, conservatrice au service des archives sonores du département audiovisuel de la BnF. Il a notamment été présenté dans le cadre du symposium NINJAL France/Japon de 2013.

Tout corpus sonore déposé à la BnF rentre dans la chaine d'archivage :

> *« Les corpus sonores déposés au département Audiovisuel, après avoir été inventoriés, cotés, restaurés si besoin et numérisés, sont ensuite intégrés à l'archive numérique, archive numérique permettant bien sûr une pérennité des documents déposés et la préservation de leurs originaux, mais aussi l'exploitation scientifique et la diffusion aux publics de ces archives sonores. Avant cette diffusion aux publics, le département accorde une grande attention au traitement documentaire des fonds déposés. Ce traitement différant légèrement, selon le fonds, bien que reposant sur une base commune de travail : l'écoute, l'identification, la description et l'indexation. On pourrait distinguer, même si une continuité de traitement est assurée entre eux, fonds anciens et fonds plus récents. La distinction ne se faisant pas sur leurs natures, la continuité des domaines d'études sujets de ces corpus étant valorisée et leur mise en relation privilégiée. Les fonds anciens, historiques ou bien dont le ou les producteurs sont disparus (Archives de la Parole, Musée de la Parole et du Geste, etc.). Ces fonds enregistrés peuvent présenter des états lacunaires quant aux supports, ou bien encore aux archives papiers, photographiques… ayant pu être produites simultanément. Le traitement documentaire, en plus de ses aspects classiques d'identification, d'écoute et de description, essaie autant que possible, après un travail d'enquête et de recoupements (avec ses propres fonds ou ceux d'autres institutions), de redonner à ces fonds une histoire et une analyse les plus complètes, en valorisant tous les aspects qui se doivent de l'être, qu'ils soient techniques, scientifiques ou bien encore historiques mais aussi en les mettant en relation intellectuelle avec d'autres fonds. Pour les fonds plus récents et bénéficiant d'un dépôt 'complet' (archives matérielles, métadonnées...) que les institutions/personnes productrices, en activité, ont décidé de déposer dans nos services, le traitement documentaire et l'exposition sont choisis d'un commun accord, suivis d'un travail collaboratif afin de proposer une valorisation optimale du travail réalisé dans ces corpus, cas du corpus déposé par le Laboratoire Ligérien de Linguistique (LLL), l'Enquête Socio-Linguistique d'Orléans (ESLO), réalisée au début des années 70, enquête visant à donner un état du français parlé avant 1980 à Orléans et la région de l'Orléanais (Viault 2014[59]).*

Le travail de la BnF va donc compléter le travail d'exploitation scientifique. Outre, la pérennisation dans le respect des meilleures pratiques mondiales, le corpus sera ainsi mis en relation avec d'autres fonds à partir d'un traitement archivistique et accessible à tout public

[59] Viault A., Inédit, Notes pour une communication aux journées NINJAL, 18 novembre 2013.

puisque décrit dans le catalogue dédié à ces fonds d'archives, BnF-Archives et manuscrits (BAM). *[Linked open data]*

> *« Ce catalogue utilise le format connu de l'EAD (Encoded Archival Description), format basé sur le langage XML-EAD qui permet de structurer des descriptions de documents d'archives, cette description reposant sur une norme internationale de description mais aussi une norme nationale, permettant une harmonisation des pratiques et de la description des données au travers des différentes institutions dépositaires de ce genre de fonds. Le traitement de ce corpus dans BnF-Archives et manuscrits (BAM) se concrétise donc par un travail documentaire sous forme d'instrument de recherche, accessible pour tous et permettant :*
>
> *a - la signalisation des corpus sonores et des informations administratives, techniques et de gestion*
>
> *b - la description de ces corpus : identifiants (cotes...) - localisation, institution/personne productrice, contexte de production, mentions de responsabilité intellectuelle, caractéristiques matérielles (supports, état de conservation, techniques de lectures, volumétrie...), lacunes ou contraintes matérielles, contenus, modalités d'accès et de reproduction éventuellement, substituts numériques et plateforme de consultation (en salle ou Gallica selon droits)...*
>
> *c - des informations complémentaires : documents en relations, données bibliographiques, etc.*
>
> *d - des indexations (référentiels nationaux : autorités BnF et RAMEAU) et une mise en relation via ces indexations à d'autres ressources documentaires, et ce, à tous les niveaux de la description (exemple d'un composant détaillé).*
>
> *e - un moteur de recherche permettant une recherche multiple interne dans l'IR (index, recherche par mot) et externe (recherche croisée dans les IR, index et recherche par mot) du département de l'Audiovisuel et de l'ensemble des départements thématiques entrant leur fonds sur BnF-Archives et manuscrits.*
>
> *Toutes ces informations sont organisées via une description hiérarchisée, reposant sur un principe d'héritage des informations et une circulation dans*

> *et hors de l'Instrument de Recherche permise par l'utilisation de liens actifs» (Viault 2014[60]).*

En plus du signalement et de l'exposition du corpus, le travail de description et d'indexation à l'aide d'un instrument de recherche tourné en partie vers l'interopérabilité des données et des métadonnées offre une ouverture vers d'autres ressources.

Ces ressources sont celles des autres catalogues de la BnF mais aussi celles d'autres catalogues. A un second niveau, la description mais également la structuration des données dans un cadre d'interopérabilité offre une gestion de la connaissance partagée totalement bouleversée. Nous reviendrons par la suite sur l'expérimentation d'ESLO dans le cadre du web sémantique mais précisons que les documents présents dans BAM sont accessibles via Data.bnf.fr.

Data.bnf.fr est un pivot entre les ressources documentaires, qui rassemble des données numériques et des données descriptives des différents catalogues de la BnF, basé sur de nouvelles techniques de modélisation et sur le langage RDF (Resource Description Framework), avec une exposition en "link open data", compatible avec les principes du Web sémantique.

> *Cet outil, créé depuis 2011 par la BnF dans une démarche d'adoption des outils du Web sémantique, propose donc des pages indexables et référençables par les différents moteurs de recherche, contrairement aux données et métadonnées cachées dans les bases de données non indexables de la BnF (obligeant le chercheur a forcément entrer par le portail de la BnF et y découvrir éventuellement ces données). Chaque page rassemble des liens, des services et du contenu autour d'un concept informatif : (Data permet également d'élargir la recherche sur des résultats en relation sur des sites extérieurs) (Viault 2014[61]).*

On le constate, en partant d'un simple projet de conservation du corpus ESLO, on arrive pour peu que soit réalisé le travail en partenariat avec les institutions compétentes, à une véritable transformation d'un objet dédié à la connaissance.

[60] Viault A., Inédit, Notes pour une communication aux journées NINJAL, 18 novembre 2013.
[61] Viault A., Inédit, Notes pour une communication aux journées NINJAL, 18 novembre 2013.

Conclusion : la transformation d'un objet et des pratiques

A ce stade de notre étude, nous pouvons résumer les principaux effets produits par la conversion numérique du corpus ESLO :

- La « maniabilité » d'un grand corpus sonore. Ce point pourrait être un détail purement pratique mais pour qui a tenté d'écouter 315 heures d'enregistrements répartis sur plus de 470 bandes magnétiques, le confort d'une manipulation de fichiers sonores numériques est indéniable. Avant même d'avoir recours à des outils du TAL pour une navigation dans les données, le simple fait de pouvoir écouter avec un player permettant une navigation d'extraits en extraits en quelques clics fait basculer la ressource sonore dans un autre univers. La parole peut aussi être feuilletée. Enfin, pouvoir diffuser le corpus en ligne le rend facilement maniable et partageable.

- La synchronisation son-texte
  Nous l'avons esquissé et nous le développons dans un chapitre spécifique mais la transcription d'ESLO a été conçue comme un simple outil de navigation dans le signal sonore et non comme une étape constitutive de données primaires. Toutefois il s'agit bien d'une démarche de linguiste et la transcription est la première (chronologiquement) des annotations. Il s'agit ni plus ni moins d'une information catégorisante, reliée à un segment sonore. Ceci dit, dans le cas de ce type de transcription, l'annotation est volontairement très pauvre, ayant comme simple objectif d'accompagner la donnée d'origine vers un traitement outillée de celle-ci. In fine, ce sont bien les pratiques des chercheurs qui deviennent numériques bien plus que les données.

- La description structurée des données et leur contextualisation.
  La numérisation entraine également une transformation du document en lui-même, ce qui implique une description particulière. Mais elle offre aussi une ouverture vers un nouvel objet auquel on peut relier, de manière stable, des éléments contextuels. Pour ne citer que l'exemple de la TEI, on voit que les métadonnées y sont embarquées dans l'entête.
  Au-delà d'une simple description des données par des métadonnées, l'enjeu est de fournir un accès à des données contextualisées, ce qui

transforme aussi fondamentalement la démarche linguistique fondée sur l'observation comme le souligne Laks 2010 [62] :

> « *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*[63] ».

- Le liage du corpus à la connaissance partagée.

Les données numériques naissent dans un univers très sensibilisé, par nécessité, à la normalisation et à la standardisation. Si cet état de fait requiert une prudence légitime de la part des chercheurs, il a un effet positif dans le « liage » des données entre elles et, avant d'en arriver à ce stade, dans le signalement et la mise en relation de corpus différents. Cette « connectibilité » d'objets scientifiques modifie chaque élément de cette connaissance partagée.

L'exemple de cette expérience de conservation d'ESLO dépasse donc un objectif de pérennisation d'un corpus et transforme véritablement un premier objet scientifique en un second dont les contours, les formes et le périmètre sont modifiés. Cette modification ne se réduit pas à l'objet en lui-même mais également aux points de vue qui se trouvent bouleversés par les nouvelles pratiques d'observation qui en découlent.

Ce travail a été particulièrement enrichissant mais également très lourd et il représente une part importante de mes activités de recherche de ces dix dernières années. Il s'agit d'une expérience de recherche conduite non pas directement par les données mais par la gestion de celles-ci, de la collecte à la diffusion en passant par la conservation et l'exploitation. Il en ressort un va-et-vient systématique entre opérations de gestion du projet, méthodologie, analyses des données et réflexion sur les conditions de production. Il me semble que c'est la maitrise de l'ensemble de cette chaine qui est au principe d'une activité scientifique satisfaisante.

[62] LAKS, B. (2010). « Langage et variation : pourquoi y a-t-il de la variation plutôt que rien ? ».
[63] Sur le paradoxe de l'observateur *Cf.* Labov (1975), Labov (1976). Sur le marché linguistique et les effets d'observation *Cf.* Encrevé (1976), Encrevé (1982)

3.3 ESLO2 (2004-…) [retour]

Articles et livre :

- 2008, Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO, https://halshs.archives-ouvertes.fr/halshs-01162506
- 2009, Un grand corpus oral « disponible » : le corpus d'Orléans, 1968-2012, https://halshs.archives-ouvertes.fr/halshs-01163053
- 2011, *(Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?,* https://hal.archives-ouvertes.fr/hal-01162479
- 2015, Pourquoi et comment dresser le portrait sonore d'une "grande ville"? L'exemple d'ESLO2, https://halshs.archives-ouvertes.fr/halshs-01165945

Communications orales :

- 2006, Constitution et exploitation d'un grand corpus de "données situées" Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), https://halshs.archives-ouvertes.fr/halshs-01165954
- 2008, Un grand corpus de référence du français parlé : état des lieux et perspectives, https://halshs.archives-ouvertes.fr/halshs-01165952, https://halshs.archives-ouvertes.fr/halshs-01165954
- 2009, Les Eslos, un corpus variationniste représentatif d'une communauté d'auditeurs ? https://halshs.archives-ouvertes.fr/halshs-01165949
- 2014, « ESLO : du portrait sonore à la "ville " », https://halshs.archives-ouvertes.fr/halshs-01165911

Documents :

- Site ESLO Architecture du corpus : http://eslo.huma-num.fr/index.php/pagecorpus/pagepresentationcorpus
- Portrait de territoire ESLO2 : https://www.nakala.fr/nakala/data/11280/55690730
- Affiche Eslo2 (metadata)
- Bulletin de participation ESLO2 (metadata)
- Complément au guide du transcripteur: Lexique (metadata)
- Formulaire de consentement ESLO2 (metadata)
- Formulaire locuteur ESLO2 (metadata)
- Guide du matériel ESLO2 (metadata)
- Guide du transcripteur/relecteur des ESLOs V1 (metadata)
- Guide du transcripteur/relecteur des ESLOs V2 (metadata)
- Guide du transcripteur/relecteur des ESLOs V3 (metadata)
- Guide du transcripteur/relecteur des ESLOs V4 (metadata)
- Plaquette de présentation ESLO2 (metadata)
- Procédure dépôt ESLO2 (metadata)

3.3.1 Petite épistémologie d'ESLO2 [retour]

ESLO2 est un projet de constitution et d'analyse d'un second corpus réalisé à partir d'une nouvelle enquête sociolinguistique à Orléans dans les années 2010. Ce projet est en cours de réalisation par l'équipe ESLO du Laboratoire Ligérien de Linguistique (UMR 7270, ex EA CORAL).

Dès l'origine du projet, l'objectif était double puisqu'il s'agissait (i) de réaliser cette nouvelle enquête qui devait être comparable à la première tout en prenant en compte l'évolution des cadres théoriques et de la technologie et (ii) de développer un projet de diffusion et d'exploitation scientifique d'un corpus constitué de l'ensemble des deux enquêtes.

VARILING

Ce projet a pu véritablement démarrer avec la sélection du projet VARILING (Traitements des variations linguistiques dans les corpus) dans le cadre de l'appel à projet de l'Agence Nationale pour la recherche (ANR Corpus 2006).

> *« Dès sa conception, ESLO 2 a été conçue pour préfigurer une référence dans un domaine qui, à l'échelle internationale, est en structuration et où l'adoption d'un format standardisé de collecte, de conservation, de traitement et d'analyse est confrontée à la multiplicité des développements et des normes. La certification est construite d'abord en tenant compte des pratiques en usage concernant la fabrication des corpus oraux en linguistique. A partir de l'échange entre les acteurs de la recherche, une synthèse des recommandations sera effectuée par le consensus des usagers et des experts, déterminant une conception du traitement d'ESLO 1 et de fabrication d'ESLO 2 qui permette d'exemplariser ces deux fonds pour de futurs corpus (et leurs traitements). »*

Le projet VARILING présente très clairement les orientations scientifiques du projet. Le cœur en est la variation linguistique, or celle-ci s'appréhende au sein d'un corpus d'enquête sociolinguistique dont la chaine de traitement doit être totalement maitrisée. Dès cette présentation on perçoit également la nécessité de placer le projet au sein d'un champ en cours de constitution : celui de la linguistique de corpus numérique :

> *Si ESLO 2 a une visée cumulative (accroitre la quantité de données pour assurer des comparaisons avec d'autres), l'enquête est aussi réflexive (accompagner l'enquête, le traitement et l'exploitation d'une analyse de l'expérience pour contribuer à la définition des normes). Cette conception concerne :*
> *- une prospective sur l'exhaustivité des usages avec un calcul de représentativité,*
> *- un inventaire des techniques de collecte (formats d'enregistrement et numérisation),*
> *- une politique de formation des enquêteurs et d'information des témoins afin d'intégrer dans les critères de variation celle liée à l'enquêteur avec pour projet l'organisation d'une école d'été sur ce thème,*
> *- un recueil des données concurremment à l'enrichissement en méta-données,*
> *- un codage et un catalogage anticipant les principales requêtes émergeant en linguistique, mais aussi en sociologie, en anthropologie, en histoire, en info-com…*
> *- une transcription avec alignement dans une perspective de normalisation,*
> *- un étiquetage, avec catégorisation et lemmatisation (en particulier, recherches sur les problèmes rémanents de disfluences de l'oral et de coréférence anaphorique en situation de parole spontanée),*
> *- une procédure d'anonymisation (l'identification des questions posées par l'anonymisation permettra la confection d'un vade mecum des éléments à prendre*

en considération à cette étape du travail à partir d'une recherche sur la détection des entités dénommantes (et pas seulement des entités nommées)),
- un stockage, avec archivage et indexation,
- une procédure de mise à disposition : la construction et la maintenance du site doivent assurer une libre consultation sur Internet (avec une convivialité et une ergonomie des applications, si possible dans une version multilingue),
- des données partagées : interopérabilité et protections, en liaison avec les propositions formulées dans le cadre du programme pour le catalogage et codage des corpus CatCod qui prolonge le travail de l'EPML50, « Corpus d'interaction langagière ». Les spécifications retenues seront transmises au consortium « Text Encoding Initiative » (TEI) à titre de proposition.
Au-delà, seront dessinées la mise en place du suivi (maintenance, jouvence et sécurité) et les applications. Il ne s'agit pas d'anticiper les analyses, mais de les rendre possibles.
Ce projet témoigne d'une ambition forte. Constituer en prototype, à toutes les étapes de sa réalisation, un corpus qui puisse se situer au même niveau, qualitatif et quantitatif, y compris par sa dimension patrimoniale, que les grands corpus oraux fabriqués, ou en cours de fabrication, en Europe et dans le monde.

En termes de données l'objectif est symbolique : atteindre les 10 millions de mots en combinant 5 millions de mots d'ESLO1 et 5 millions de mots d'ESLO2, soit plus de 700 heures d'enregistrement. Cet objectif n'est pas anecdotique. Le seuil des 10 millions de mots était fréquemment utilisé dans les discussions autour d'un corpus de référence du français parlé, or Claire Blanche-Benveniste qui s'est inlassablement battue pour un tel objectif, évaluait le cout de revient d'un corpus de cette taille à un euro le mot, soit 10 millions d'euros. Dans le contexte de l'époque, il était fort peu vraisemblable d'obtenir une telle somme, la question d'une autre méthode pour atteindre cet objectif est donc pertinente si on prend en compte le contexte institutionnel et budgétaire des conditions de la recherche.

Au-delà des effets de la politique de la recherche sur ses conditions de production, Le projet VARILING se situait dans la configuration théorique suivante :

- Sociolinguistique et corpus variationniste

Ce cadre est présenté dans un chapitre spécifique. On précisera néanmoins que l'objectif de VARILING est de prendre en compte les évolutions dans ce domaine au cours des quarante années qui séparent la réalisation des deux corpus. Or force est de constater que si la sociolinguistique apparaissait comme une approche très prometteuse à la fin des années soixante, elle ne s'est pas imposée massivement dans les années qui suivirent. Jusqu'aux années quatre-vingt, il y eut une montée en puissance notamment à partir des travaux d'Encrevé sur la liaison[64] que confortait l'apport de la sociologie de Pierre Bourdieu et ses

[64] ENCREVÉ, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*.

effets, entre autres, sur la linguistique[65] mais la sociolinguistique n'a pas eu l'effet qu'elle escomptait sur la linguistique comme le rappelle Bergounioux (1992:5)

> *« Il n'empêche, la linguistique de terrain, et singulièrement la sociolinguistique, semblent s'essouffler, incapables de rivaliser avec les théories les mieux constituées, ou les plus légitimées par le champ scientifique, la grammaire générative, si décriée, gardant à cet égard un prestige indiscutable. La sociolinguistique s'enseigne encore mais sans autre référence que celles qu'elle s'était données à sa fondation : Bernstein, Bourdieu, Labov pour ne citer que les contemporains, sans non plus progresser de façon significative à partir d'eux »*[66].

L'héritage scientifique est toutefois suffisamment dense pour exercer ses effets sur la réalisation d'une nouvelle enquête.

- Linguistique interactionnelle

Le second cadre théorique est celui de la linguistique interactionnelle. Issue de la pragmatique et flirtant avec la sociolinguistique, la linguistique de l'interaction est clairement un domaine qui s'est développé après le projet ESLO de la fin des années soixante. Il n'est donc pas étonnant de ne trouver que des traces un peu naïves de cette approche, notamment le souhait d'enregistrer des « données authentiques issues de conversations familières ».

ESLO1 comprend des tentatives d'enregistrement de repas (http://cocoon.huma-num.fr/exist/crdo/meta/crdo-ESLO1_REPAS_272), de conversations familiales (http://cocoon.huma-num.fr/exist/crdo/meta/crdo-ESLO1_REPAS_555), d'enregistrements « sur le vif » sur les marchés (http://cocoon.huma-num.fr/exist/crdo/meta/crdo-ESLO1_MAR_609) , de prises de contact (http://cocoon.huma-num.fr/exist/crdo/meta/crdo-ESLO1_ENTCONT_201) etc.

Outre la qualité exécrable de nombreux enregistrements, ceux-ci sont, à l'opposé des entretiens, très éloignés de toute référence théorique et de toute rigueur dans la captation et dans la définition des objectifs.

Quarante ans après, l'ethnométhodologie, l'analyse de conversations et les corpus d'enregistrements de situation « non provoquées par le chercheur » ont été considérablement développés et une discipline est apparue comme le souligne Mondada (2008:881) :

[65] BOURDIEU, P. (1982). *Ce que parler veut dire: l'économie des échanges linguistiques*.
[66] BERGOUNIOUX, G. (1992). « Les enquêtes de terrain en France ».

> *« La linguistique interactionnelle est un paradigme récent, qui a émergé comme tel durant les années 90, tout en reposant sur les acquis de l'analyse conversationnelle, apparue dans les années 60. Elle répond de manière spécifique à ces exigences, en développant un projet systématique d'étude de la langue dans l'interaction, sur la base d'enregistrements d'interactions en situation naturelle. Très présente dans la linguistique scandinave, anglo-saxonne et allemande, elle l'est encore peu dans le paysage de la linguistique française, tout en commençant à s'y développer, comme en témoignent les articles présents dans cette section »*[67].

Il s'agit pour VARILING de ne pas négliger l'apport de la linguistique interactionnelle pour les corpus d'enquêtes sociolinguistiques.

- La linguistique de corpus

Il y a incontestablement une différence très forte dans la conception de la linguistique de corpus entre les années 1960 et 2000. Certes, comme le souligne B. Laks, la linguistique de corpus a toujours existé ; elle n'est pas apparue avec le développement de l'informatique :

> « *A jeter un regard rétrospectif sur nos disciplines se découvre ainsi un vaste panorama dans lequel l'utilisation systématique de descriptions ordonnées sous la forme de base de données et de corpus de référence constitue une pratique ancienne et récurrente, bien établie méthodologiquement et pratiquement en linguistique et en philologie. Il n'y a donc d'anachronisme que de façade à annexer ces pratiques à une linguistique de corpus apparue en tant que courant de recherche en sciences du langage de façon bien plus récente* » (Laks, 2008b [68]).

Toutefois on ne peut contester l'impact de l'ingénierie informatique et du traitement automatique des langues dans les pratiques de linguistique de corpus (Habert, Nazarenko et Salem 1997[69]). Dans le cas d'ESLO où le premier corpus n'a connu aucun traitement informatisé de la part des responsables du projet et où la perspective d'un traitement automatique n'a pas été ne serait-ce qu'esquissé dans les objectifs du projet, le gap est particulièrement sensible.

[67] MONDADA, L. (2008). « Contributions de la linguistique interactionnelle ».
[68] LAKS, B. (2008). « Pour une phonologie de corpus ».
[69] HABERT, B., NAZARENKO, A., & SALEM, A. (1997). « Les linguistiques de corpus / Benoît Habert, Adeline Nazarenko, André Salem ».

C'est donc dans une perspective à la fois comparatiste et cumulative que le projet d'une nouvelle enquête et d'un nouveau corpus (ESLO2) s'inscrit. Face à l'ampleur de la tâche de réalisation d'un très grand corpus oral de français, le projet est tout d'abord orienté vers les aspects méthodologiques sans que ceux-ci soient séparés des cadres théoriques et des perspectives d'analyse. On retrouve cette orientation dans la description des différentes phases de traitement du corpus dans le projet initial (VARILING:48) :

Phase 1 : Exhaustivité, représentativité, proportionnalité
Phase 2 : Techniques de collecte : formats d'enregistrement et numérisation
Phase 3 : Formation des enquêteurs et information des témoins
Phase 4 : Recueil des données
Phase 5 : Codage et catalogage
Phase 6 : Transcription et alignement
Phase 7 : Étiquetage, catégorisation et lemmatisation
Phase 8 : Anonymisation
Phase 9 : Stockage, archivage et indexation
Phase 10 : Mise à disposition : construction du site et traduction
Phase 11 : Données partagées : interopérabilité et protections
Phase 12 : Applications et développements
Phase 13 : Mise en place du suivi (maintenance, jouvence et sécurité) et applications

Le projet ESLO2

Le projet ESLO2 est donc né dans le cadre des activités d'un laboratoire qu'il a accompagné et souvent précédé dans son développement. A Orléans, la première esquisse d'une équipe de recherche en linguistique était fondée sur une collaboration entre des collègues de l'Ecole Normale d'Orléans (actuelle ESPE) et deux enseignants-chercheurs de l'UFR Lettres, Langues et Sciences Humaines. Baptisée « Groupe de Recherche sur l'Ecole et le Langage à Orléans » (GRELO), cette unité, soutenue en son temps par le conseil scientifique de l'établissement, a commencé ses travaux par des enregistrements en milieu scolaire. Les études se sont élargies à la sociolinguistique avec le recrutement de nouveaux collègues (justifié par la création d'une filière FLE puis l'ouverture d'un département de sciences du langage) et le rapatriement des enregistrements d'ESLO cependant que le GRELO se transformait en Centre Orléanais de Recherches en Acoustique et Linguistique (CORAL) reconnu comme UPRES-JE (Unité Propre de Recherche à l'Enseignement Supérieur – Jeune Equipe) par le Ministère à compter du 1er janvier 1996. Lors du contrat 2000-2003, la proposition de constitution d'un grand laboratoire fédérant les lettres et les langues de l'Université d'Orléans a été refusée par le Ministère et le quadriennal a dû être assumé dans une structure qui avait repris le périmètre du CORAL sans labellisation. C'est sous ce sigle, désormais décliné comme Centre Orléanais de Recherches en Anthropologie et Linguistique pour tenir compte d'un rapprochement avec l'IRD autour d'une description des langues de Guyane, que le laboratoire a été reconnu comme UPRES-EA (Equipe d'Accueil) en 2004. Sur

proposition de collègues linguistes de Tours, une réunion de leur équipe « Langage et Représentations » avec le CORAL a donné naissance au LLL en 2008 et, par association avec le Département de l'Audiovisuel de la BnF et le CNRS, le LLL est devenu l'UMR 7270 depuis le 1er janvier 2012.

Il doit donc une part de son existence à l'intérêt du directeur du laboratoire et à son intérêt pour une linguistique variationniste fondée sur l'enquête, Gabriel Bergounioux (Bergounioux 1992 :19) :

> « *Oui, et aussi cela : dominée par une linguistique qui ne décrit plus rien d'observé, et même, parfois, peut-être, plus rien d'observable, la linguistique de terrain, sous des noms divers, oppose des interrogations et ménage des crises : théorie du discours, réalisme phonologique, construction du sens, statut de la syntaxe de l'oral, autant de questions qu'une théorie classique, fût-elle générative, peine à dénouer. Si l'on reprenait un exercice qui faisait rire hier, vers 68 justement, de commenter les exemples des grammaires pour calculer la valeur des théories qui s'en justifient, on pourrait s'apercevoir en quoi une linguistique de l'oral nous a fait, à défaut d'être plus savants, moins crédules. Pas d'avancée en linguistique sans ce questionnement de l'oral et pas d'oral sans enquête, sans corpus, sans témoin. Si ce numéro constitue un bilan, il ne serait pas mal venu qu'il soit, aussi, un appel* »[70].

puis à mon arrivée dans l'équipe, à la formation que j'ai suivie auprès de Pierre Encrevé au sein du groupe de linguistique variationniste de l'EHESS dans les années 90, à ma décision de prendre en charge la numérisation d'ESLO et, dans la continuité, la réalisation d'ESLO2, à réactiver un projet tombé en déshérence.

Il n'est pas anodin qu'ESLO2 soit élaboré dans la continuité des travaux de numérisation et de mise à disposition d'ESLO. Les objectifs de conservation et d'accès aux données se retrouvent au cœur d'une démarche sociolinguistique d'observation et d'analyse de la variation linguistique. Avec un peu de recul, nous pouvons mettre en avant les points les plus significatifs du projet des ESLOs :

- En premier lieu, l'objectif d'analyse variationniste qui s'appuie sur une prise en compte des conditions de production des données. Cet objectif sera abordé dans un exemple d'analyse sur la liaison ; il représente le véritable centre du travail autour duquel s'organise l'ensemble des autres éléments de constitution et de traitement du corpus.

[70] BERGOUNIOUX, G. (1992). « Les enquêtes de terrain en France ».

- La question de la représentativité d'un corpus variationniste. Nous retrouverons cette question dans le travail réalisé sur la conception de l'architecture d'ESLO2.

- L'enjeu et l'impact des questions juridiques et éthiques. Cette question se concrétisera dans la démarche de collecte, la gestion des données personnelles dans le corpus et la gestion des droits de diffusion.

- Le rôle du catalogage et de la contextualisation des données. Objectif déterminant dès le projet ESLO1, celui-ci prend tout son sens dans la réalisation d'un entrepôt de données conservé et diffusé selon les pratiques actuelles de l'archivage scientifique.

- La transcription. Première opération pour les corpus oraux de linguiste, celle-ci sera présentée sous une forme très différente pour la reprise d'ESLO1 et la réalisation d'ESLO2.

Ce sont ces points significatifs qui structurent les chapitres suivants.

3.3.2 Préalable à l'élaboration de l'architecture d'ESLO2 [retour]

Le corpus ESLO2 répond prioritairement à un double objectif : permettre une comparaison avec ESLO1 et être représentatif du français parlé dans les années 2010. L'architecture du corpus ESLO2 est donc une réponse construite pour assurer la représentativité des données.

De façon générale, poser la question de la représentativité, c'est procéder avant tout à un inventaire des causes de variation. On les distribuera entre trois catégories :

- celles qui sont liées à des propriétés intrinsèques des locuteurs, avec une ventilation par âge, par sexe, par CSP, par trajectoire sociale et par origine géographique, sur le modèle de ce qui avait été fait avec ESLO1 et qui bénéficie depuis des apports cumulés de l'anthropologie et de la sociologie (on intègrera, en face de la CSP, la définition donnée par le témoin de son statut social) ;

- celles qui relèvent des situations de discours et des niveaux de langue exploités, y compris les actes performatifs et la construction des identités relationnelles, la définition du cadre, l'organisation des tours de parole, la construction du point de vue, saisis dans la variété des situations de collecte et des modes de relation entre enquêteur et enquêté ;

- celles qui articulent formats cognitifs (type de tâche à accomplir : récit, description d'un itinéraire, argumentation, plaisanterie, demande d'assentiment…) et exploitation des ressources linguistiques (figures du discours, opérations syntaxiques, choix lexical, deixis et procédures anaphoriques, références spatio-temporelles…).

Ces trois versants, distingués pour la commodité de l'exploitation, sont corrélés. Ils se caractérisent par leur caractère obligé (tout discours en tant qu'il est oralisé est assignable à un locuteur, un contexte de production et une tâche cognitivo-linguistique). Ils doivent de ce fait constituer la charpente d'un corpus dont l'architecture entend répondre à un objectif de représentativité. Il convient toutefois d'interroger cette notion et celle de « corpus de référence » avant de passer à la description de l'architecture du corpus.

Un corpus de référence ? *[corpus-référence]*

Si le terme de « corpus de référence » a été très vite utilisé au début du projet ESLO2, il a été aussi assez vite abandonné devant les malentendus que provoquait cette notion. La première ambigüité tient à l'usage de cette expression pour deux notions différentes, l'une venant du TAL pour qui le corpus de référence est un jeu de données permettant d'entrainer et de tester des outils d'annotations automatiques dans le cadre d'apprentissage automatique notamment, l'autre en provenance de la linguistique descriptive pour laquelle le corpus de référence doit contenir suffisamment de données pour assurer la représentativité des « faits de langue ».

Ainsi la question de la représentativité du corpus ESLO2 a d'abord donné lieu à une réflexion sur la notion de corpus de référence en linguistique. Il en ressort la volonté de construire un corpus qui contienne des informations explicites sur la valeur de données qui soient représentatives des pratiques linguistiques au sein d'une *communauté d'auditeurs*.

Une communauté d'auditeurs

Dans la même perspective qu'ESLO1, ESLO2 s'appuie sur la représentativité d'une communauté d'auditeur appréhendée au sein d'une « ville », ce qui nécessitait de définir les périmètres de l'objet ville et les « axes » de variation linguistique : diatopique, diastratique et diaphasique.

La variation diatopique a été appréhendée de la même manière qu'ESLO1 : elle ne répond résolument pas à un objectif de représentativité du français dans le monde, ni même en France, ni même en France métropolitaine.

> *« Mais pour des raisons d'ordre théorique et pratique, pour écarter aussi des variables incontrôlables, il a été décidé que le projet d'enquête serait réalisé dans le cadre d'une ville donnée. Orléans, communauté urbaine moyenne (commune de 100 000 habitants en 1968), exempte de caractères dialectaux très marqués, cité historique mais en même temps ville en plein essor, démographique, économique social et culturel (taux de croissance annuel de population de 3 pour 1000, nombreuses industries nouvelles, université récente, développement du secteur tertiaire), capitale régionale échappant suffisamment à l'attraction de Paris, a paru offrir l'homogénéité indispensable et la variété recherchée »*. [Blanc & Biggs 1971, p. 16]

En revanche, le corpus a pour objectif de prendre en compte la langue telle qu'elle est reçue et donc produite au sein d'une communauté d'échanges linguistiques. Les notions de variations diastratique et diatopique mériteraient d'être rediscutées car elles laissent supposer que la variation est structurée en axes indépendants or ceux-ci sont systématiquement présents. On ne peut par exemple opposer simplement deux entretiens sous prétexte que les locuteurs correspondent à deux catégories sociales différentes, de même qu'on ne peut opposer l'entretien à une autre situation d'échange comme celle d'un repas. Dans les deux cas, l'identité sociale des locuteurs-interactants dépend avant tout d'un marché linguistique où ces deux axes sont intriqués. L'appréhension d'un marché linguistique est bel et bien ce que cherchaient à construire les auteurs d'ESLO1 :

> « *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]

ce qui conduit à dresser le portrait sonore d'une ville

> « *à l'intérieur de laquelle reconstruire, à un moment précis, la dynamique des formes linguistiques simultanément présentes dans une cité assez vaste pour que la variation y soit accusée et perpétuée à travers des réseaux linguistiques d'échanges autonomes, et assez restreinte pour que n'importe quel membre de cette communauté linguistique ait dû interférer dans les circuits de communication des autres groupes*" [Bergounioux et al. 1992 : 79].

La ville *[observatoire]*

Le choix d'ESLO1 était très simple concernant l'aspect géographique. Il s'agissait de sélectionner une ville de taille moyenne dans une région où l'accent est perçu comme standard. Il était très novateur sociologiquement en prenant pour objet une communauté d'auditeurs. La ville sera d'ailleurs, à partir des années 1970, un terrain privilégié pour la sociolinguistique urbaine.

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

1. Appréhender L'objet « ville »…

(…) C'est sur ces constats que se déploie le champ de la sociolinguistique urbaine qui, pour reprendre Gasquet-Cyrus (2002 : 55), prend quatre directions majeures. Une partie des travaux se concentrent sur « les changements observés dans la distribution des langues (transmission, véhicularisation) en milieu urbain ». C'est ce qui apparait dans les travaux de Calvet (1994, 2000, entre autres). On y constate la façon dont les langues en contact dans les échanges urbains s'auto- et hétéro-régulent. D'autres visent « les effets de la ville sur les formes linguistiques », ou comment la « polyphonie urbaine », pour reprendre Lamizet, contribue à faire évoluer la langue et génère de nouvelles formes. D'autres encore tentent de cerner les représentations à travers les discours pour éclairer les « identités urbaines ». Enfin, une quatrième orientation est notable, bien qu'elle cristallise en fait les problématiques qui articulent les trois autres. Elle concerne tous les travaux s'intéressant à la banlieue et aux pratiques des jeunes. Dans tous les cas, il est bien question de reconnaitre la validité de l'objet ville relativement à des pratiques langagières qui le définissent, peut-être mieux que n'importe quel autre critère objectivable : « dans la ville, l'expérience des rapports sociaux ne peut être que verbale » (Lamizet, 2002 : 15).

Le sociolinguiste a ainsi naturellement à dire sur la ville et ce qui s'y passe, à la condition d'y voir un espace d'hétérogénéité et de dynamisme par essence, qui n'existe pas a priori mais se dessine dans les discours. Déjà, en 1994, Calvet mettait en garde : « la sociolinguistique urbaine ne peut pas se contenter d'étudier des situations urbaines, elle doit dégager ce que ces situations ont de spécifique, et donc construire une approche spécifique de ces situations.» (1994 : 15). On n'observe donc pas la ville comme l'on observe un autre territoire. Toute tentative de recueil de données illustrant quelque chose de la ville, devrait impliquer une réflexion d'ordre méthodologique tenant compte de la redéfinition constante du territoire exploré, parce qu'il n'est pas donné d'emblée mais se précise, voire se détermine, au regard de ce que disent lesdites données. On soulève là un paradoxe incontournable, qui place le chercheur dans une posture où il a à assumer de ne pas être celui qui assigne les catégories (Mondada 2000, 2002). S'intéresser à la ville implique que soient, à chaque étape du travail, mis en regard les positions théoriques, les principes méthodologiques et les données (Gadet & Guerin, 2012).

Déjà dans ESLO1 l'objectif était de définir la ville à partir des pratiques :

> *« Ecouter parler des Français de divers milieux qui en répondant à certaines questions ou en s'exprimant librement sur certains thèmes, révèleraient leurs présuppositions, leur expérience vécue et la façon dont ils la jugent (…) les besoins définis pour la recherche socio-linguistique, nous ont menés à la notion d'un "portrait sonore d'une ville". En effet nous aurions la double possibilité d'observer des attitudes et des expériences ressenties individuellement, et d'étudier à travers des expériences communes, la vie d'une communauté et un cadre institutionnel réel »*. [Blanc & Biggs 1971:22]

Cette référence à une sociologie des pratiques sera particulièrement développée dans ESLO2 :

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

La ville est restée l'élément central d'un projet présenté comme la réalisation du « portrait sonore de la ville ». Le questionnaire a notamment été revu afin de réorienter l'entretien semi-directif vers une conversation moins formelle dont le thème est la vie de chaque « témoin » dans la ville d'Orléans. La discussion est ainsi orientée vers les pratiques du locuteur au sein de la ville. Outre un cadre prédéfini selon une souplesse assumée par les enquêteurs, ces conversations permettent de recueillir des indicateurs sociologiques particulièrement pertinents. Enfin l'architecture même du corpus ESLO2 a été totalement repensée pour appréhender ce que l'on a coutume d'appeler la variation diaphasique. S'il est maintenant

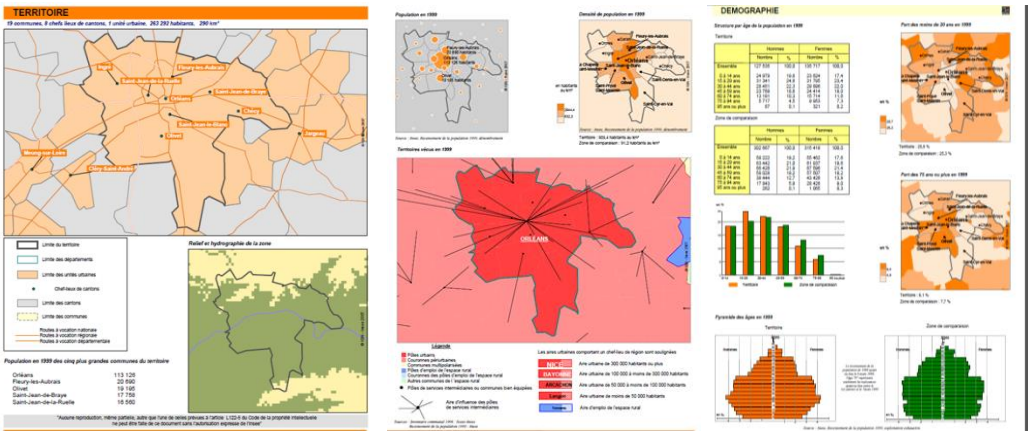
classique de différencier des productions linguistiques sur un axe paroles publiques (formel) paroles privées (informel), il est rare que des travaux plus ambitieux, intégrant par exemple les paramètres relatifs à la distance/proximité communicationnelle (Koch & Oesterreicher, 2001), soient pris en compte de manière significative dans les grands corpus. L'architecture d'ESLO2 repose quant à elle sur une catégorisation de « situations » d'enregistrements très diversifiés qui souhaite rendre compte des apports de la linguistique interactionnelle et de la sociolinguistique dans la description des pratiques linguistiques.

Si le périmètre ville n'était pas questionné dans ESLO1, il le sera dans ESLO2 d'une part pour prendre en compte l'apport d'indicateurs socioéconomiques disponibles à partir des années quatre-vingt, et d'autre part pour intégrer l'évolution de la structure des villes vers les agglomérations. Ainsi dans ESLO2 la ville d'Orléans sera considérée comme intégrant les communes de l'agglomération dans un périmètre élargi.

La description socioéconomique s'appuie sur des données de la statistique publique et d'analyses fournies par l'INSEE à partir du dernier recensement en vigueur (1999 en l'occurrence pour des chiffres publiés en 2007) et synthétisées dans un document à l'aide de 19 tableaux et d'autant de cartes : «Portrait de territoires ».

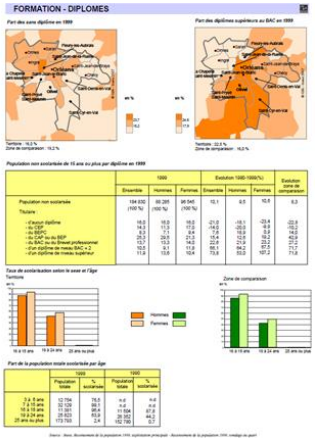
Je ne peux ici présenter l'ensemble du travail effectué mais voici quelques extraits significatifs pour décrire ce pré-portrait socioéconomique[71] sur lequel s'appuie le projet d'un portrait linguistique.

- Zone du territoire et population

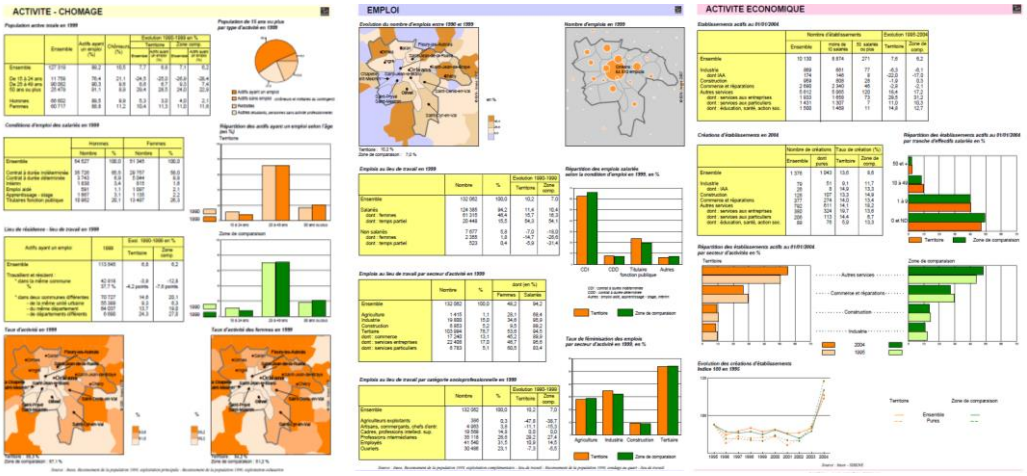


- Formation et diplômes :

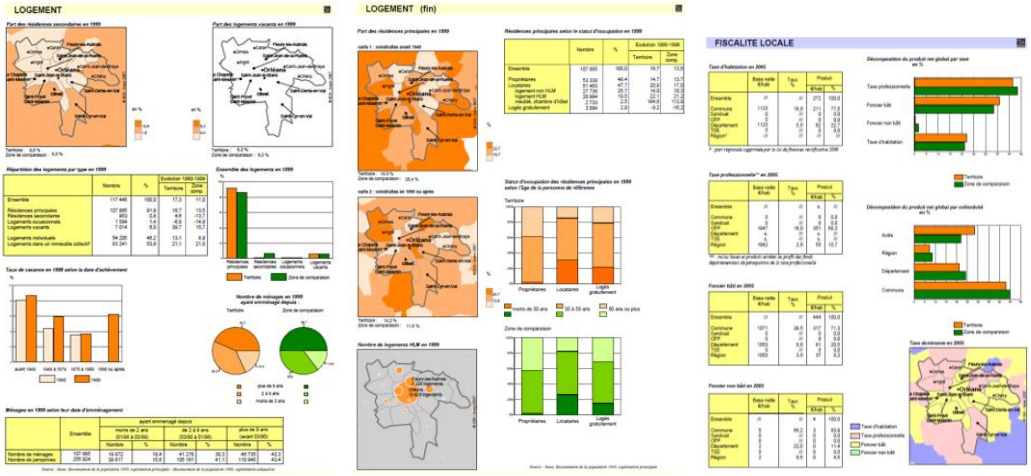
[71] Portrait de territoire, INSEE 2007



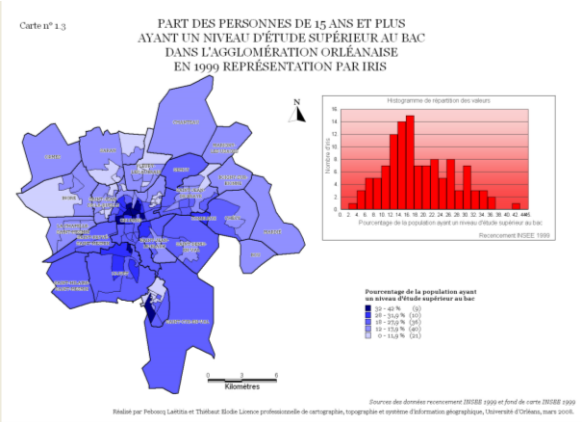
- Activité Chômage et emploi

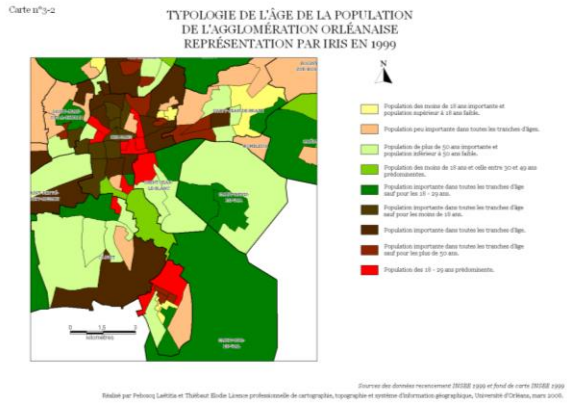


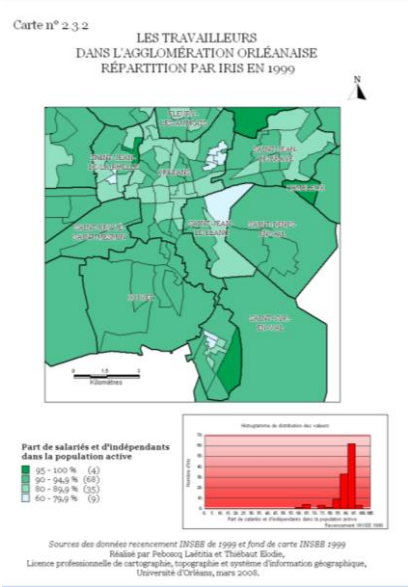
- Logement et Fiscalité locale

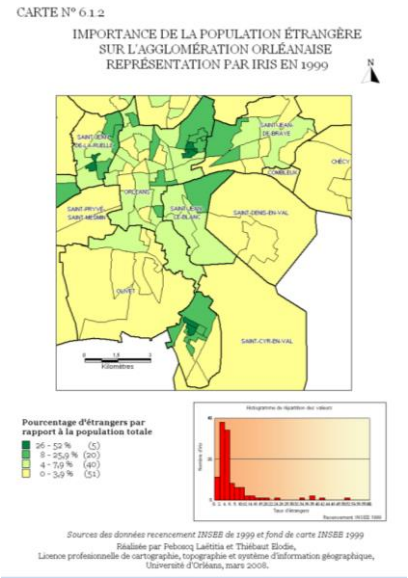


Il faut noter que dans le cadre du programme ESLO2, un projet a été réalisé avec l'équipe de géographie de l'Université d'Orléans, afin de réaliser des fonds de cartes nécessaires au traitement des données croisées (socio-économiques et linguistiques) et un « atlas social » à partir des données fournies par l'INSEE. Ce sous-projet est l'un des exemples des exigences induites par la réalisation d'une enquête sociolinguistique rigoureuse qui se propose de décrire la variation linguistique. Voici quelques extraits significatifs de cet atlas social :









Ce travail d'agrégation et d'analyses des données socioéconomiques a pour objectif d'outiller la partie sociologique de l'enquête. Il appelle trois remarques :

- Premièrement, la sociolinguistique, et notamment la sociolinguistique « urbaine » si cette dénomination est pertinente, doit prendre en compte et intégrer l'outillage fourni par

les nombreuses données socioéconomiques disponibles. Force est de constater que ce n'est que rarement le cas et que la plupart des études s'appuient sur un « savoir non explicite » du linguiste qui « connaitrait » son terrain. Dans le cadre d'ESLO2, ces connaissances sont exploitées pour réaliser le panel des entretiens.

- Deuxièmement, les sciences humaines et sociales doivent considérer la masse et l'hétérogénéité des données disponibles pour des travaux et des études qui ont de plus en plus souvent recours à des approches trans- ou multidisciplinaire. Le développement des Humanités numériques et l'apport d'infrastructures comme la TGIR Huma-Num pour les corpus et la TGIR Progedo pour les données quantitatives constituent une opportunité pour conduire à bonne fin un vaste chantier à peine ouvert.

- Troisièmement, la présentation minimaliste des données sous forme de cartes nous incite à plaider en faveur d'une restitution des travaux en « data visualisation » qui semble aujourd'hui la méthode la plus efficace pour accompagner et guider le chercheur dans des analyses conduites à partir de données multiples et massives.

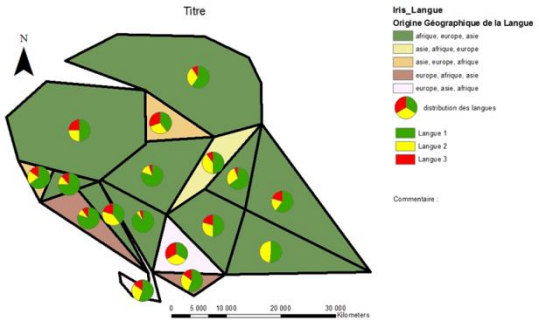
Pour conclure cette partie dans une perspective directement liée au projet ESLO, je donnerai l'exemple du travail exploratoire mené dans le cadre d'une collaboration entre linguistes et géographes à l'université d'Orléans afin de représenter à l'aide de cartes les données issues de l'enquête sur la transmission familiale des langues réalisées en 1999 par l'INED[72]. L'objectif ici est de prendre en compte les langues en contact avec le français parlé à Orléans[73].

Les différentes cartes souhaitées sont :

- Carte représentant les langues parlées entre l'enfant et ses frères et sœurs
- Carte représentant les langues parlées entre l'enfant et sa mère
- Carte représentant les langues parlées entre l'enfant et son père
- Carte représentant les autres langues parlées dans la famille
- Carte représentant les langues parlées entre l'enfant et ses ami(e)s
- Carte représentant les langues parlées entre le père et ses ami(e)s
- Carte représentant les langues parlées entre la mère et ses ami(e)s
- Carte représentant le nombre de langues parlées dans l'Iris

Exemple de carte :

[72] HÉRAN François, FILHON Alexandra, DEPREZ Christine, " La dynamique des langues en France au fil du XX^{e} siècle ", *Population et Sociétés*, n°376, février 2002.
[73] Projet Langues en contact à Orléans (Responsable Jean-Louis Rougé).



Ce travail initié en 2008 a été suspendu en attendant de disposer de premières pistes d'analyses linguistiques pertinentes exploitables par des spécialistes de datavizualisation et de cartographie.

3.3.3 Architecture d'ESLO2 [retour]

Le module « entretiens » et la stratification sociale

La réalisation d'ESLO1 et son exploitation ont été principalement centrées autour des entretiens. Nous l'avons souligné, c'est sur ce point que la méthodologie d'ESLO1 fut la plus rigoureuse. Elle est en totale cohérence avec l'approche de la linguistique variationniste prônée par Labov à la même époque (Labov 1975)[74] pour qui le fait linguistique est indissociable des faits sociaux, ce qui implique la prise en compte d'une stratification sociale qu'il définit dans *Sociolinguistique*[75] :

> *« La stratification sociale est le produit de la différenciation et de l'évaluation sociale. Le terme n'implique aucunement l'existence de classes sociales ou de castes spécifiques, mais signifie simplement, que le fonctionnement normal de la société a produit des différences systématiques entre certaines institutions ou certaines personnes, qui ont été hiérarchisées d'un commun accord sur une échelle de statut ou de prestige* ». (Labov 1976:96)

ESLO2 est donc bâti à partir d'un module d'entretiens semi directifs qui a pour objectif, outre la comparaison directe avec ESLO1, de recueillir des enregistrements de locuteurs diversifiés selon les critères classiques de la sociologie. Le panel constitué s'appuie sur les critères suivants : âge, sexe et catégorie socio professionnelle. Cette catégorisation sommaire ne peut être considérée comme satisfaisante et la trame de l'entretien a été conçue pour obtenir des informations sur les pratiques sociales et culturelles des témoins

[74] LABOV, W. (1975). *What is a linguistic fact ?*
[75] LABOV, W., & ENCREVE, P. (1976). *Sociolinguistique*.

dans la ville afin de permettre une critérisation beaucoup plus fine après l'analyse de l'entretien. Pour cette phase du travail de constitution du corpus, le thème des pratiques citadines et la méthodologie de l'entretien « non violent » seront particulièrement utiles.

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

3. Observer et rendre compte des activités dans la ville

La ville, sans doute plus que n'importe quel autre terrain, voit émerger des activités langagières difficilement catégorisables a priori, tant les interactions sont multiples. Y émergent des phénomènes, effets des contacts de cultures, de langues, de formes de langue, qui sont difficilement prévisibles compte tenu de l'inévitable porosité des frontières du territoire (voir supra). On peut cependant considérer que chaque interaction est caractérisable selon la combinaison de paramètres suivante : degrés de planification du discours, d'interactivité, de distance sociale entre les interactants, de convergence et de formalité du cadre.

La ville est un terrain privilégié pour observer l'imbrication des contraintes et influences individuelles et collectives. « La polyphonie urbaine est une médiation, en ce qu'elle institue une dialectique entre le singulier et le collectif : chacune des voix qui participent à la polyphonie urbaine a son identité propre, distincte et reconnaissable, mais, dans le même temps, l'ensemble de ces voix se retrouvent dans un projet collectif, dans un système collectif qui exprime une identité d'ensemble. » (Lamizet, 2007 : 21). Dessiner le portrait sonore d'une ville implique dès lors de s'intéresser aux activités langagières en ce qu'elles s'organisent selon des normes collectives (macro) et des normes individuelles (micro).

Contrairement à ESLO1, si ce module prend une place particulière et centrale, il n'est néanmoins qu'un des éléments de l'architecture du corpus.

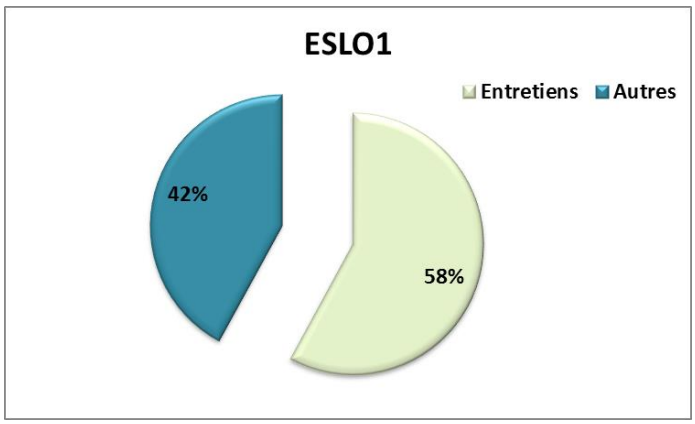
L'architecture va considérablement évoluer dans le cadre du corpus ESLO2 afin de prendre en compte l'avancée méthodologique et théorique réalisée entre 1968 et 2008. D'une part l'évolution technologique a une forte incidence sur la collecte des corpus oraux. Si les auteurs d'ESLO1 se félicitaient de disposer d'un 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 matutinale à la soirée entre amis en passant par l'activité professionnelle et les conversations familiales.

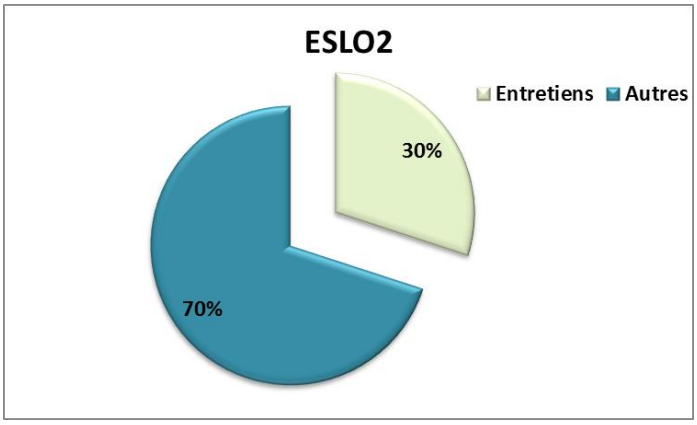
Cette évolution technologique s'accompagne d'un engouement fort pour la captation d'enregistrements diversifiés dans des situations non provoquées par le chercheur selon les objectifs de l'analyse de conversations.

L'objectif de dresser un portrait sonore ne peut donc se résumer à la collecte d'entretiens selon un échantillonnage sociologique. Il convient parallèlement d'élaborer une architecture de corpus qui permette de rendre compte de la diversité des situations de production et d'audition. Force est de constater que la méthodologie d'ESLO1 était balbutiante à cet égard. Si les entretiens ont été réalisés avec beaucoup de rigueur, les autres types d'enregistrements sont souvent de mauvaise qualité, voire inexploitables, et correspondent à des objectifs peu maitrisés. L'enregistrement d'une même personne dans diverses situations s'est réduit 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, représentatives des pratiques d'une communauté.

C'est toute l'architecture du corpus qui s'en trouve modifiée pour qu'apparaissent dans leur diversité des situations de productions linguistiques qui soient répertoriées au sein d'un marché linguistique plus global.

Evolution de la place des entretiens d'ESLO1 à ESLO2 :





3.3.4 Catégorisation des modules et architecture générale

[http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus] [retour]

Le travail sur l'architecture du corpus recoupe différentes approches qui convergent vers les effets de catégorisation des modules. Ceux-ci peuvent en effet renvoyer à des « situations de communication », « genre », « styles », « registre », à des objectifs différents, des méthodologies variées où à une recherche concernant l'un des axes de variations décrits.

Cette analyse est présente dans le chapitre consacré à la notion de corpus et de « corpus de référence ». On peut recenser quelques approches disponibles :

Le genre (Biber 1988[76],1999[77]) :

23 genres majeurs en tout, 6 pour l'oral.
- Conversations en face-à-face
- Conversations par téléphone
- Débats et entrevues en public
- Émissions de radio et télévision
- Discours non préparés
- Discours planifiés

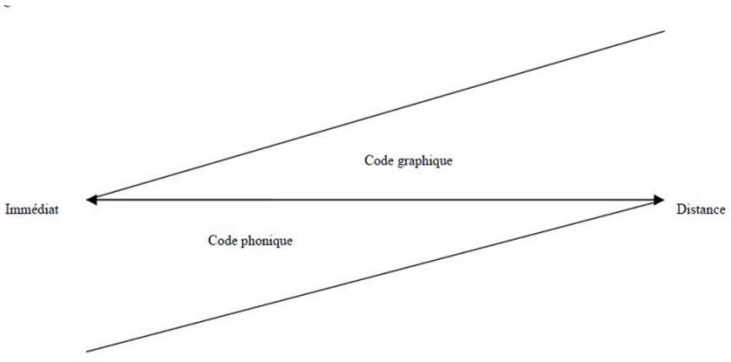
Le style en sociolinguistique :

"*Il est possible d'aligner tous les styles sur une seule dimension, que mesure le degré d'attention portée au discours*" (Labov 1972; 1976: 288)
"*Linguistic choices rarely index social categories directly; rather, they index attitudes, stances, activities that are in turn associated with categories of people. It is this indirect nature of the relation between variables and categories that allows variation to be a resource not simply for the indexing of place in the social matrix but for the construction of new places and of nuanced social meanings*" [Eckert 2005 http://www.stanford.edu/~eckert/csofp.html]

Le registre :

« variété isolable d'une langue employée dans des situations sociales définies » (Ferguson 1982)

Les travaux de Koch & Oesterreicher (2001) tentent une analyse plus fine qui repose sur un continuum.



Gadet reprend cette conception dans son ouvrage de 1983 sur la variation sociale en français (Gadet 1983:53[78]).

[76] BIBER, D. (1988). *Variation across speech and writing*.
[77] QUIRK, R. (1999). *Longman grammar of spoken and written English*.

Proximité communicationnelle	Distance communicationnelle
Communication privée	Communication publique
Interlocuteur intime	Interlocuteur inconnu
Emotionnalité forte	Emotionnalité faible
Ancrage actionnel et situationnel	Détachement actionnel et situationnel
Ancrage référentiel dans la situation	Détachement référentiel dans la situation
Co-présence spatio-temporelle	Séparation spatio-temporelle
Coopération communicative intense	Coopération communicative minime
Dialogue	Monologue
Communication spontanée	Communication préparée
Liberté thématique	Fixation thématique

De fait, l'architecture d'un corpus ne peut se résumer au pourcentage des genres, styles ou situations représentés sans que soit interrogée la pertinence de cette catégorisation au sein d'une structure globale.

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

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.

Les travaux en linguistique de corpus apportent peu de réponse actuellement à ces questions. D'une part les initiatives de standardisation de ces catégories sont balbutiantes, que ce soit en TEI ou dans les formats de description archivistique, d'autre part les projets de corpus se résument souvent à une répartition schématique. C'est le cas du corpus DELIC qui présente ainsi la structure du CRFP en 2004 :

> *a) la parole privée : sous la forme d'un entretien sollicité spécifiquement dans le cadre de l'enquête, cette situation de parole renvoie à deux types de productions : le récit de vie (dont le contenu peut varier : récit d'un voyage, d'une expérience, souvenirs d'enfance, …), ou la présentation d'un « savoir-faire » professionnel ou autre.*

[78] GADET, F. (2003). *La variation sociale en français*.

b) la parole professionnelle : entretiens également sollicités spécifiquement, mais dans lesquels les locuteurs ont été enregistrés dans l'exercice de leur fonction ou quand ils parlent de leur profession sur leur lieu de travail.

c) la parole publique : cette situation se distingue des deux autres par le fait que les intervenants s'expriment toujours en présence d'un public ; elle comporte une partie d'entretiens sollicités, le reste étant constitué d'émissions radiophoniques (actualités, interview, table ronde, tribune téléphonique, etc.), de cours et conférences, de réunions politique ou associative (conseil municipal, discussion syndicale, comité de quartier, etc.), et de quelques situations plus spécifiques (visite de musée, dégustation de vins, etc.). (DELIC 2004)

ou plus récemment du projet PFC :

PFC est un projet qui vise à décrire la prononciation du français dans sa diversité géographique, sociale et stylistique. (2002:5)
Le but de l'enquête est de rassembler un échantillon de variétés de français et la procédure détaillée ci-dessous nous permet d'avoir accès à la variation individuelle. L'ensemble des enregistrements permet en effet une étude de plusieurs registres chez le même locuteur :
- dans les entretiens, un français soutenu et parfois un français familier
- dans les dialogues, un français familier
- dans la lecture, un français très soutenu (PFC Bulletin N°1 2002:9)

Ou encore du projet CIEL-F dans la version 2008 :

« Le nombre de situations devant être limité et se retrouver dans différents pays, on en a retenu quatre :
1) un cours magistral ou discours public
2) des conversations libres spontanées,
3) des entretiens « épilinguistiques »
4) des enregistrements sur les marchés

Dès ESLO1, le projet était nettement orienté vers une prise en compte de la diversité des types d'enregistrements. Cette approche n'est pas formalisée mais elle se concrétise de différentes manières. Si on se réfère au texte de présentation de Blanc & Biggs (Blanc & Biggs 1971:19) on peut relever 6 types d'enregistrement « *exemple de langages plus spontanés recueillis dans des situations variées* » dont la description relève de critères, non explicites, linguistiques et extralinguistiques :

- Entretiens,
- Témoins dans d'autres situations,

- Micros cachés,
- Entretiens de personnalités,
- Tables-rondes et débats, conférences,
- Entretiens au CMPP.

Si on se réfère aux choix de catalogage réalisé par l'équipe, on comptabilise huit catégories (Lonergan 1974:1) :

1. Entretiens,
2. Reprise de contacts,
3. Témoins dans d'autres situations,
4. Communications téléphoniques,
5. Interviews sur mesure de personnalités,
6. Conférences débats,
7. Enregistrements divers comportant des témoins inconnus,
8. Interviews au CMPP

Le catalogue précise les attentes concernant chacun des groupes à partir de critères linguistiques « comparaison entre interview et contexte moins structuré », « forme de dialogue médiée » et/ou sociologiques « témoignages de personnalités publiques parlant de leurs rôles et activités ».

Dans le même catalogue, au moment de présenter le codage des bandes, on relève seulement 7 catégories (Lonergan 1974:2) :

- 001-200 : groupe 1
- 201-300 : groupe 2 et 3
- 301-400 : groupe 4
- 401-500 : groupe 5
- 501-600 : groupe 6
- 601-700 : groupe 7
- 701-800 : groupe 8

L'écart tient au regroupement des catégories 2 et 3 sans que cette réduction soit justifiée.

Ajoutons, comme cela a été mentionné, que les opérations de numérisation et d'intégration du catalogue dans le process des ESLOs ont conduit à une autre catégorisation dans ESLO2, sous la forme de comme « modules », une notion au demeurant assez lâche.

Pour ESLO2, un travail a été mené conjointement sur la collecte d'enregistrements dans des situations variées et sur la définition de modules à partir de concepts sociolinguistiques opératoires. Ainsi le corpus ESLO2 présente une grande diversité de situations d'enregistrements afin de répondre à un objectif de couverture étendue de la diversité des pratiques linguistiques dans la vie quotidienne. A

l'origine, deux grands principes (linguistique et sociologique) ont conduit à dessiner une première architecture en fonction de cinq axes :

1. La planification du discours
2. La structure de l'interaction
3. La distance sociale entre les locuteurs
4. La convergence discursive et interactionnelle
5. Le cadre, plus ou moins formel

Mais il a paru surtout important de concevoir le corpus ESLO comme un réservoir de données qui ne vise pas une représentativité exhaustive mais qui contienne suffisamment d'informations contextuelles pour permettre des analyses contrastives. Ainsi les nombreux modules d'ESLO2 donnent lieu à une fiche qui présente ses caractéristiques et incluent une conservation des documents afin de permettre une contextualisation scientifique (protocoles, carnets de terrains, description des objectifs).

Voici quelques exemples de ces fiches (http://eslo.huma-num.fr/index.php/pagecorpus/pagepresentationcorpus) :

Fiche de module : Entretien ESLO2	
Catégorie	Entretien
Corpus	ESLO2
Référence	ESLO2_ENT_XXXX
Description	Discussion en face-à-face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien
Protocole	Prise de contact préalable : présentation du projet et prise de rendez-vous. Enregistrements réalisés chez les locuteurs témoins
Objectifs	Dresser des portraits sociologiques de locuteurs orléanais à partir d'entretiens qui abordent leur trajectoire personnelle et leurs pratiques situées dans la ville (journée de travail, vie de quartier, sorties culturelles, loisirs, etc.). Constituer un corpus d'entretiens semi-directifs sur la base d'un panel échantillonné (âge, sexe, catégorie socio-professionnelle)
Enquêteurs	Chercheurs du Laboratoire Ligérien de Linguistique
Enregistrements disponibles	78
Enregistrements en cours de traitement	6
Enregistrements visés	150
Conditions d'accès	Accès public
Documentation disponible	Panel - Trame d'entretien - Autorisations - Guides de transcription (V1 à

	V4)
Fiche de module : Média ESLO2	
Catégorie	Média
Corpus	ESLO2
Référence	ESLO2_MEDIA_XXXX
Description	Enregistrement d'émissions radiophoniques (journaux et autres émissions)
Protocole	
Objectifs	Parole médiatique locale
Enquêteurs	Chercheurs du Laboratoire Ligérien de Linguistique
Enregistrements disponibles	1
Enregistrements en cours de traitement	16
Enregistrements visés	50
Conditions d'accès	Accès public
Documentation disponible	Guides de transcription (V3 et V4)

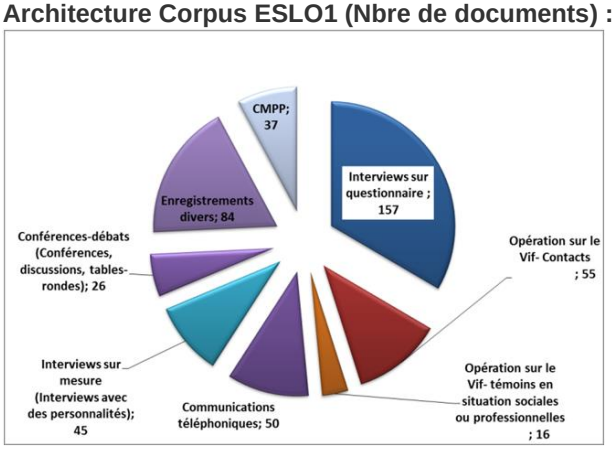
Fiche de module : Boulangerie ESLO2	
Catégorie	Boulangerie
Corpus	ESLO2
Référence	ESLO2_BOUL_XXXX
Description	Enregistrements effectués dans toutes les boulangeries d'Orléans
Protocole	Enregistrements, dispositifs non visible. Captation d'une séquence d'interaction entre un(e) vendeur(/euse) et un client(e)
Objectifs	Interactions non provoquées par le chercheur et hors d'un contexte d'enregistrement avec un dispositif visible
Enquêteurs	Etudiantes en 3e année de licence de Sciences du Langage de l'Université d'Orléans
Enregistrements disponibles	1
Enregistrements en cours de traitement	97
Enregistrements visés	98
Conditions d'accès	Accès restreint sous convention
Documentation disponible	Guides de transcription (V3 et V4)

Fiche de module : Livre pour enfants ESLO2	
Catégorie	Livre pour enfant
Corpus	ESLO2
Référence	ESLO2_LPE_XXXX
Description	Lecture de livres faite par un adulte pour un (ou plusieurs) enfant(s)
Protocole	Lecture d'un livre à un (des enfants) dans une situation habituelle (lecture familiale, lecture scolaire ou en centre de loisirs)
Objectifs	Lecture en situation non provoquée par le chercheur, rôle de la norme et

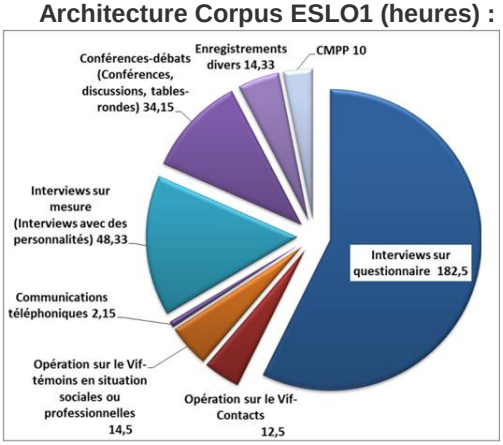
	de la transmission
Enquêteurs	Etudiantes en 3e année de licence de Sciences du Langage de l'Université d'Orléans
Enregistrements disponibles	1
Enregistrements en cours de traitement	33
Enregistrements visés	50
Conditions d'accès	Accès public
Documentation disponible	Guides de transcription (V3 et V4)

Les schémas suivants présentent en nombre de documents et en nombre d'heures0 la différence entre le corpus ESLO1 réalisé et le corpus ESLO2 tel qu'il a été conçu :

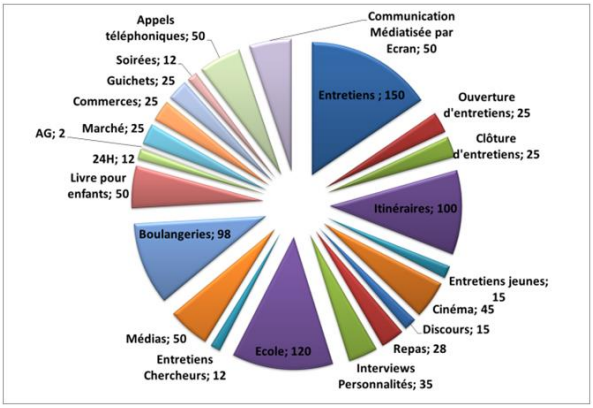
Architecture Corpus ESLO1 (Nbre de documents) :



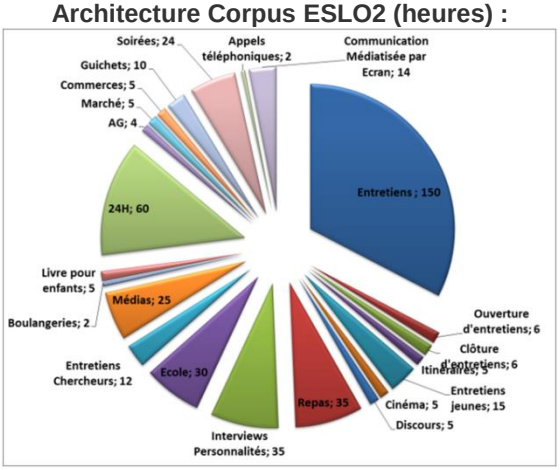
Architecture Corpus ESLO1 (heures) :



Architecture Corpus ESLO2 (Nbre de documents) :



Architecture Corpus ESLO2 (heures) :

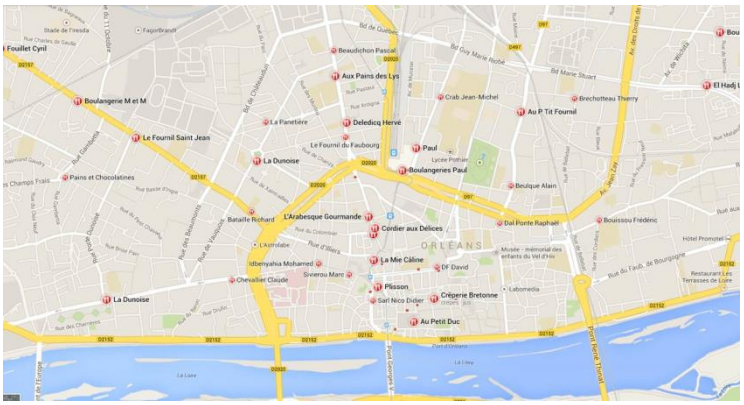


Focus sur deux modules ESLO2

Une présentation de chacun des modules serait fastidieuse dans le cadre de ce travail, néanmoins un focus sur deux d'entre eux permet de prendre la mesure du travail de terrain que requiert une collecte de ce type.

- Module Boulangerie

Ce module a été réalisé avec le concours des étudiants de sciences du langage dans le cadre du cours « Terrain Enquête Corpus ». L'objectif était de collecter un enregistrement « sur le vif » pour reprendre la terminologie d'ESLO1 ou en termes de « situation non provoquée par le chercheur » selon l'analyse de conversation. Un protocole a été établi afin d'enregistrer la même interaction dans toutes les boulangeries de la ville d'Orléans (55 répertoriées selon les pages jaunes de l'annuaire » : une séquence complète entre un(e) client(e) et un(e) vendeur (euse).



Le module « Boulangerie » comprend donc :

- un protocole,
- un guide technique,
- un guide juridique,
- un guide de transcription,
- un guide de codage et de catalogage,
- 153 enregistrements avec leurs métadonnées,
- 153 transcriptions,
- 15 carnets de terrain.

- Module Ecole

Le module école a été réalisé comme mémoire de Master recherche de Pauline Philardeau : « Les variations de la langue dans l'enseignement du français langue maternelle » co-encadré par Emmanuelle Guerin (Guerin 2015). Pauline Philardeau a été accueillie pendant plusieurs mois dans une école de l'agglomération orléanaise. A partir d'un travail sur les différentes situations catégorisables (situations de classe, entretiens, jeux dans la cour,

discussions lors de repas, réunions, etc.), de l'élaboration d'un cadre théorique sur la variation en milieu scolaire (Guerin 2012, 2015) et d'un important travail de terrain (réunions avec l'équipe enseignante, relations avec les institutions, les parents et les élèves) le corpus s'est concrétisé avec le même type de documents :

- un protocole,
- un guide technique (enregistrement en HF dans la cour d'école, enregistrements en classe, enregistrements en déplacements),
- un guide juridique (consentement de publics sensibles),
- un guide de transcription (enregistrements avec parfois 35 locuteurs),
- un guide de codage et catalogage,
- 98 enregistrements,
- 98 transcriptions,
- 1 carnet de terrain et un mémoire de recherche.

Ces deux exemples donnent un aperçu sur l'ampleur du projet et la méthode qui consiste à lier systématiquement travail de terrain, cadre théorique et méthodologique et analyses linguistiques dans un mouvement de réentrée systématique d'une approche réflexive sur la constitution et l'exploitation des données.

Etat de réalisation du corpus ESLO2

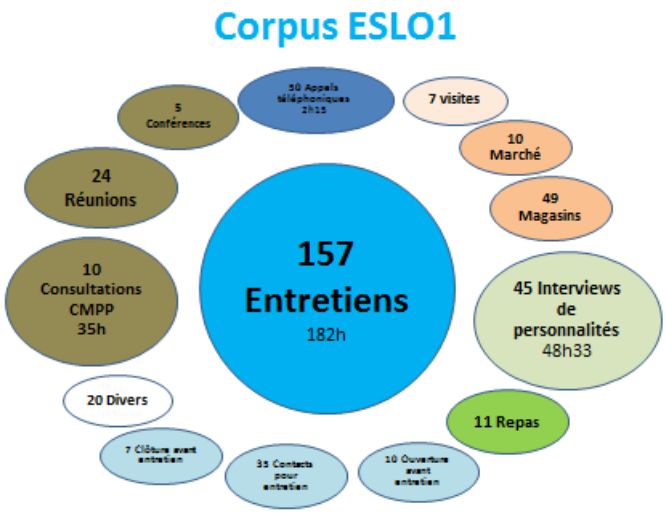
Au 1er juillet 2015 le corpus ESLO2 contient 548 enregistrements disponibles sur le site ce qui représente plus de 218 heures.

Modules	Nbre d'enregistrements	Enregistrements en ligne A+B+C	Durée enregistrements en ligne	Durée (heures)
Entretiens	85	78	85	94
Diachronie	7	7	16,5	16,5
AG	2	0		5
24H	5	1	1,6	10
Ecole	103	1		29
Livre pour enfant	43	1	0,1	3,8
Cerlico	0	0		
Personnalité	3	1	1,3	3,5
Hand	2	0		2,5
Entretiens jeunes 18-25	9	9	7	7
Boulangerie	52	1	1,6	1,5
Itinéraires	92	92	6	6
Cinéma	45	45	3	3
Conférences	6	6	4	4
Discours du 8 mai 2012	3	3	0,3	0,3
Repas	29	1	1,2	19,5
Rumeurs	38	0		2,5
Entretiens chercheurs	7	1	1,1	7
Media	17	1	0,1	3

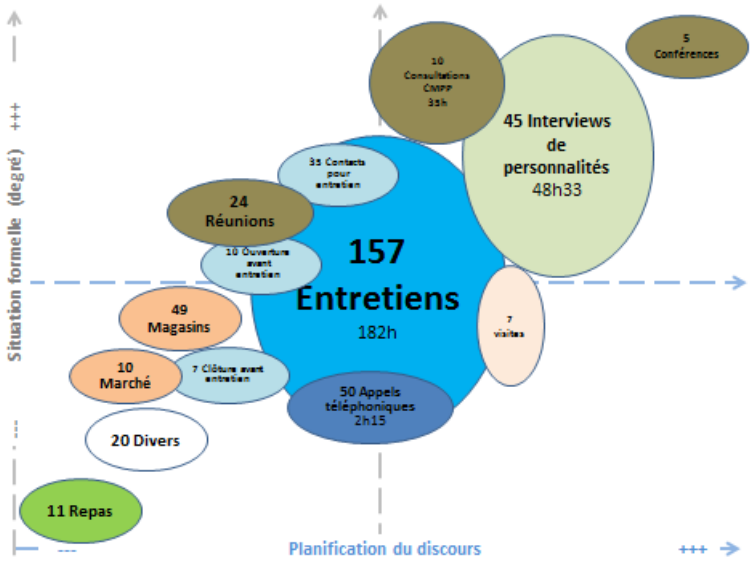
3.3.5 Catégorisation et visualisation de l'architecture [retour]

Le travail sur la catégorisation des modules ESLO2 repose sur une réflexion qui souhaite dépasser l'opposition élémentaire entre parole publique et parole privée.

A l'évidence, le corpus ESLO1 visait déjà une perspective différente dont nous pouvons visualiser l'architecture :



En première approche, nous avons catégorisé et représenté ces situations sur un graphique à partir de deux axes : le degré de formalité sociale de l'interaction et le degré de planification du discours :



Cette tentative est à la fois prometteuse et insatisfaisante. Nous butons notamment sur le manque de cadrage théorique des auteurs du corpus ESLO1.

Pour le corpus ESLO2, nous avons tenté une catégorisation dynamique qui s'appuie sur une définition pré- et post-enregistrement. Ainsi les différents modules d'ESLO2 peuvent se décrire selon un degré de prégnance suivant cinq axes :

- Degré de planification du discours (en opposant le registre « spontané » de la conversation ordinaire à celui de conférences ou le discours est écrit).

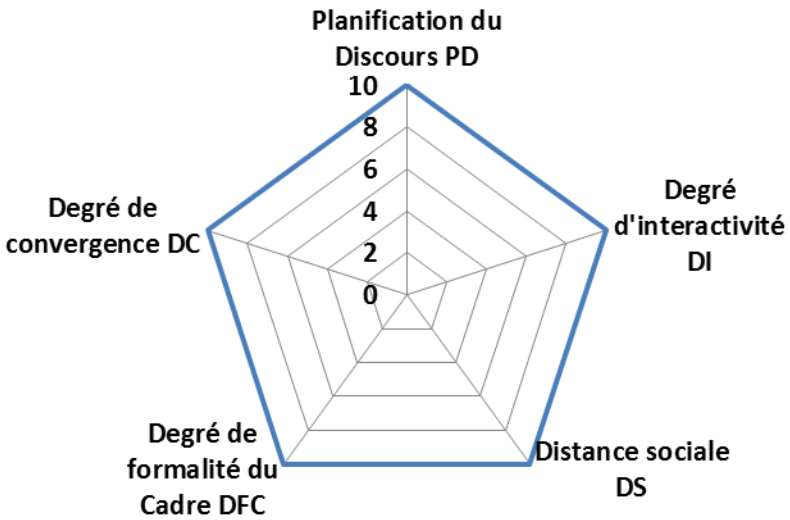
- Degré d'interactivité (du monologue au dialogue, y compris les autres conversations relevant d'un travail conséquent d'interaction).

- Degré de distance sociale entre les interactants (à partir des critères traditionnels de la sociologie : âge, sexe, niveau d'études, profession).

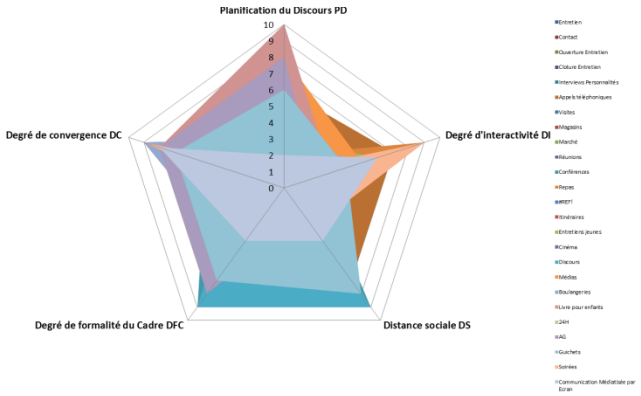
- Degré de convergence (de la polémique au consensus).

- 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).

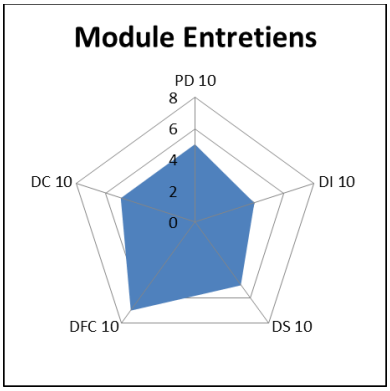
Chacun de ces critères a été évalué sur une échelle de 0 à 10 et le module peut être visualisé selon la forme obtenue par un graphique en radar :

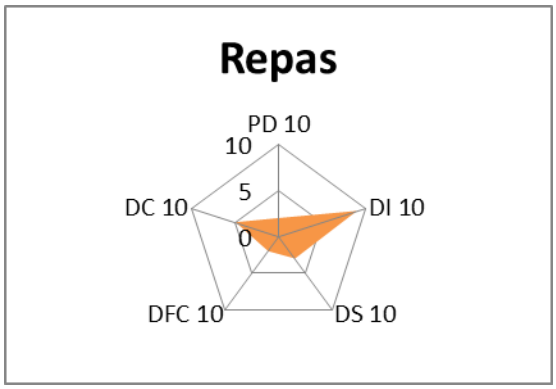


Ce qui, si l'on additionne l'ensemble des modules prévus couvrirait le spectre suivant :



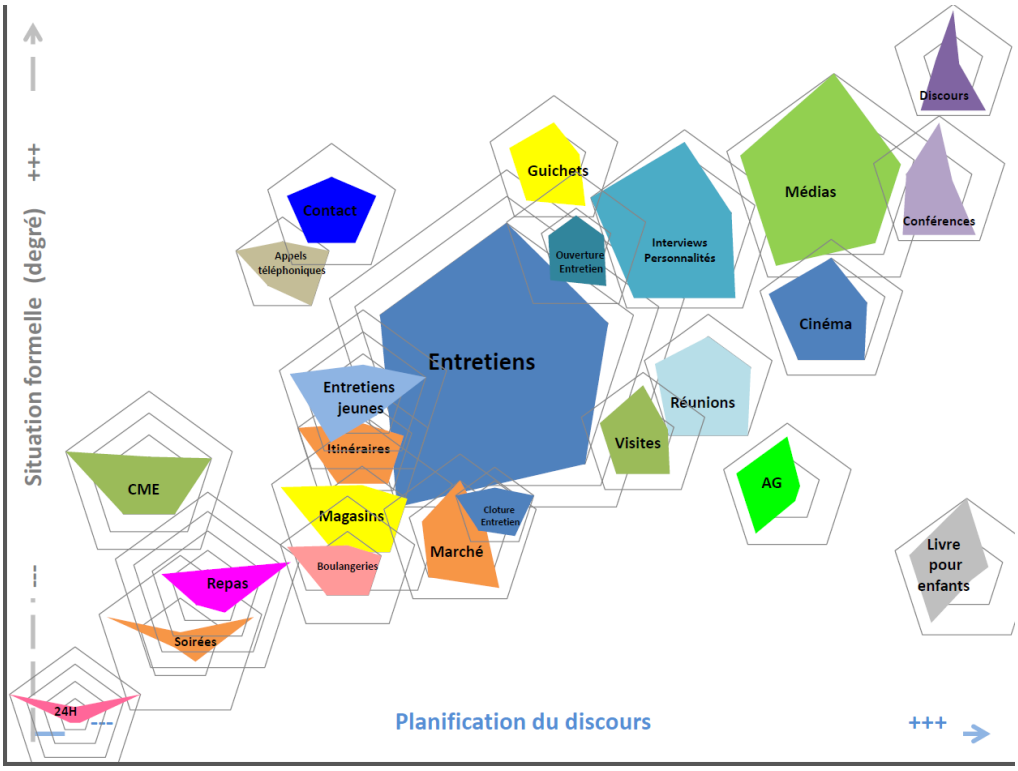
Chaque module disposant de sa propre représentation de description :





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.



Un autre intérêt de cette démarche est de permettre une catégorisation à plusieurs niveaux. Dans un article de 2015 (Baude & Guerin 2015)[79], nous proposons une catégorisation en trois étapes : (i) par module selon les objectifs recherchés, (ii) par enregistrement en prenant en compte les sonnées contextuelles et (iii) par enregistrement après analyse des données. L'analyse proposée dans l'article repose sur deux entretiens. Au niveau macro, l'entretien est caractérisé de la façon suivante :

Paramètres au niveau macro :	**E**
Degré de planification du discours (PD)	8
Degré d'interactivité (I)	4
Degré de distance sociale entre les interactants (DS)	6
Degré de convergence (C)	5
Degré de formalité du cadre (FC)	7

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

[79] Et communication au colloque Métropoles https://halshs.archives-ouvertes.fr/halshs-01165945

Par définition, un entretien suppose une certaine planification du discours et le contexte d'enregistrement à des fins scientifiques ne peut que la renforcer. Pour autant, le protocole des entretiens ESLO2, bien qu'il repose sur un scénario préconçu, incite les interviewers à s'en imprégner suffisamment pour s'en détacher et éviter un schéma discursif de type question lue/réponse stricte. De fait, nous proposons le degré 8 pour évaluer la planification du discours pour l'ensemble des entretiens.

Concernant l'interactivité, elle est inhérente à l'entretien (par opposition à une conférence, par exemple). Cependant, nous n'atteignons pas le degré 0, étant donné qu'elle est, a priori, contrôlée par l'interviewer. C'est ce qui motive le degré 4, qui rend compte de la nécessaire interactivité de l'entretien en se situant au-dessous de la moyenne mais intègre le déséquilibre des positions, en s'élevant à 4.

Concernant la distance sociale, elle est minimalement de 5 puisque les entretiens sont menés par des chercheurs auprès de non-chercheurs (l'échantillon couvre toutes les catégories socioprofessionnelles et tous les niveaux d'études). Pour autant, on n'envisage pas un degré proche de 10 car certaines informations générales d'ordre socio-démographique, laissent supposer une certaine proximité : chercheur et informateur sont adultes, habitant (travaillant) dans la région d'Orléans.

Nous évaluons le degré de convergence à 5 puisque d'une part l'objectif partagé est bien de réussir l'entretien en répondant aux questions en apportant des informations personnelles sur la vie de l'informateur, et que d'autre part le chercheur poursuit un but qui est moins explicite : obtenir un matériau dédié à l'analyse linguistique. Sur ce dernier point la convergence peut être très variable, l'informateur pouvant utiliser des stratégies pour valoriser (ou ne pas dévaloriser) son capital linguistique alors que le chercheur tentera de l'orienter vers des productions non contrôlées.

Enfin, nous évaluons le degré de formalité du cadre à 7 puisque l'entretien scientifique pose d'emblée un cadre particulièrement formel, ce qui implique d'aller au-delà de la moyenne. Néanmoins, l'équipe d'ESLO2 a entrepris un lourd travail méthodologique sur ses techniques d'enquêtes dans le but de créer un cadre le moins formel possible. Ainsi, les documents de présentation du projet « les Orléanais ont la parole », le mode d'approche, les entretiens qui se tiennent au domicile de l'informateur, la posture de l'interviewer requise et l'ensemble du protocole, sont autant de facteurs visant à atténuer le formalisme de la situation, ce qui implique que nous avons fait le choix de ne pas aller au-delà de 7.

Au niveau micro, l'évaluation des différents critères diffèrent selon les deux exemples (notons que je suis l'intervieweur-chercheur de l'exemple 2 (E2) :

Paramètres au niveau micro :	E1	E2
Degré de planification du discours (PD)	5	7
Degré d'interactivité (I)	7	5
Degré de distance sociale entre les interactants (DS)	3	7
Degré de convergence (C)	7	3
Degré de formalité du cadre (FC)	2	7

« Rendre compte des pratiques langagières dans la ville : Les ESLOs », (Baude & Guerin 2015)

Notre premier exemple d'entretien (E1) extrait du corpus ESLO concerne un enregistrement fait par un membre de l'équipe (Pauline, master recherche, 23 ans). Elle s'entretient avec une amie également âgée de 23 ans, animatrice dans le domaine périscolaire.

Notre second exemple (E2) concerne un entretien mené par un autre membre du projet (Olivier, enseignant-chercheur, 44 ans), avec un jeune homme (Nathan, ouvrier dans le bâtiment, 23 ans). Informateur et chercheur n'ont ici aucune familiarité.

Enfin l'analyse du contenu des entretiens permet une nouvelle évaluation de chaque axe. L'article présente différents éléments analysés pour chaque critère. Nous pouvons retenir à titre d'exemple particulièrement significatif dans le cadre d'un corpus permettant la comparaison d'une collection de faits, l'usage de « ouais VS oui » que j'utilise avec une variabilité particulièrement stable si l'on prend en compte les différences de statut social de l'interlocuteur dans les différents entretiens que j'ai menés :

Réf. entretien	E2	1004	1005	1014	1016	1055	1083	1272
Nbr "ouais"	327	9	4	36	26	223	222	379
Métier	Ouvrier	Institutrice	Sans	Ingénieur	Cadre	Ouvrier	Chauffeur	Sans
Niv. d'études	BTS	bac+5	Bac+5	bac+5	bac + 3	CAP	CAP	sans

Plus de 200 « ouais » en moyenne quand le témoin interviewé a un niveau d'étude faible et dix fois moins pour un bac + 5. Force est de constater que le phénomène est parfaitement corrélé à une situation sociale différente selon les interviews.

L'ensemble de l'évaluation de différents critères issus de l'analyse des entretiens permet de montrer une forte différence entre ces deux situations :

Paramètres au niveau du contenu des données :	E1	E2
Degré de planification du discours (PD)	4	3

Degré d'interactivité (I)	8	7
Degré de distance sociale entre les interactants (DS)	2	9
Degré de convergence (C)	8	3
Degré de formalité du cadre (FC)	1	5

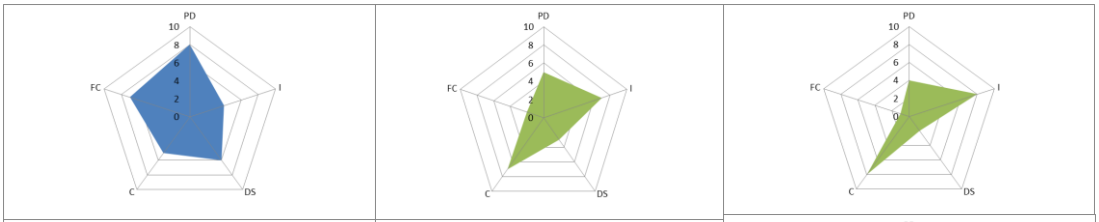
Cette démarche est en quelque sorte héritée des perspectives tracées par Alix Mullineaux si l'on s'en rapporte à ce qu'explique l'un des auteurs d'ESLO1 :

> *(...) si on veut établir des corrélations variables entre les paramètres sociaux et les écarts linguistiques – entre locuteurs d'abord, mais aussi entre les divers cadres utilisés par chaque usager de la langue – il faut un système de classement des témoins; L'échelle socio-professionnelle de l'INSEE, admirable qu'elle est pour le recensement sociologique classique, ne tient malheureusement pas suffisamment compte des paramètres culturels et du contexte familial – facteurs dont l'importance primordiale pour toute étude sociolinguistique trouve d'amples démonstrations chez Bernstein et Labov, entre autres. Nous avons pu surmonter cet obstacle grâce à l'intervention d'une psychosociologue de Birbeck College, Melle Alix Mullineaux, qui a accepté d'élaborer une nouvelle échelle de classement de la société française. Cette échelle (...) réinterprète les données de l'INSEE en y intégrant des indices de niveau culturel et de mobilité sociale relevés lors des interviews (...)* [Ross 72 p 141].

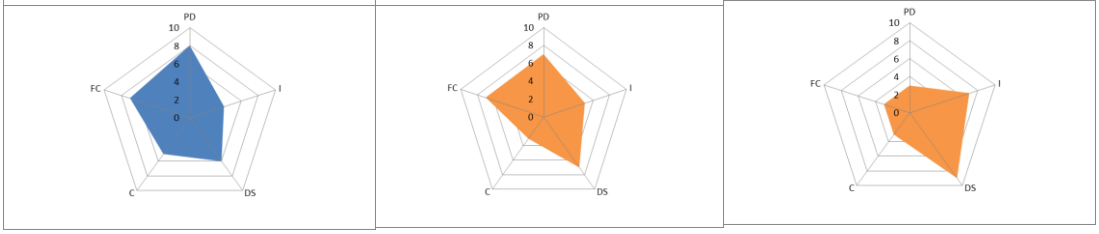
Alix Mullineaux ne mènera pas à terme son projet d'échelle AM dans cette version enrichie de l'analyse des entretiens mais l'idée est bien là.

Globalement l'évaluation des deux exemples confirme à la fois la stabilité d'un même « genre » et la nécessité d'un raffinement de la catégorisation de celui-ci à partir d'éléments contextuels et linguistiques.

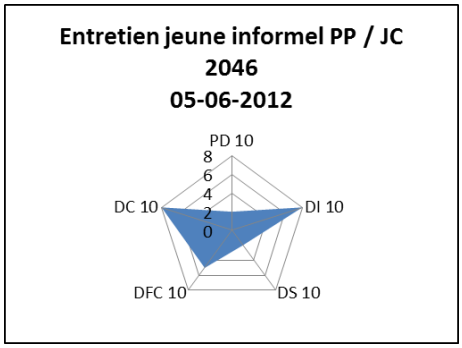
E1

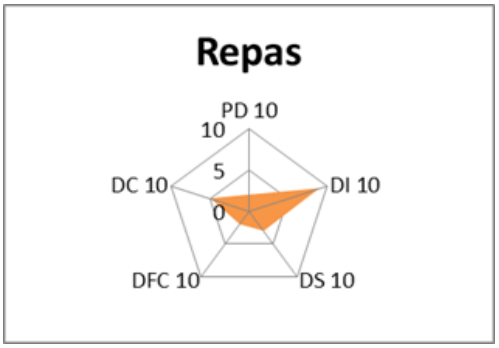


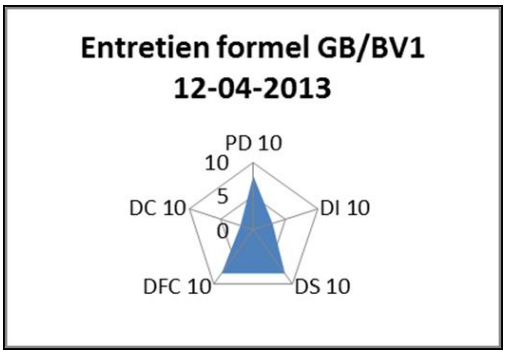
E2

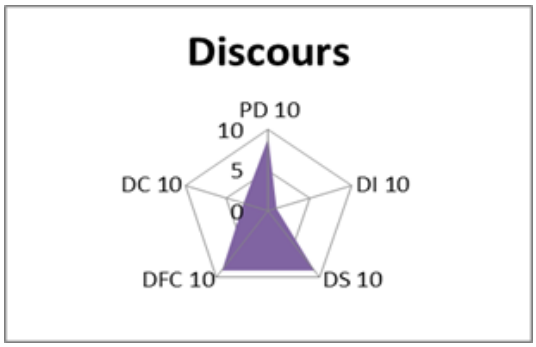


Cette finesse de description et les effets sur la relativisation des choix de catégorisation sont à rapprocher des recherches sur la méthodologie d'entretien que les sociologues ont mis à jour depuis une quarantaine d'années (Bourdieu 68[80], 93[81]). Ainsi la comparaison de deux entretiens situés aux extrêmes permet de les rapprocher de deux modules normalement catégorisés comme fortement opposés en genre :









Outre l'apport d'une démarche réflexive sur l'ensemble de la chaine de constitution et d'analyse du corpus, on peut d'ores et déjà préciser que l'évaluation des paramètres et son mode de visualisation ont notamment l'avantage de faire clairement apparaitre l'intérêt de ne négliger ni les métadonnées de niveau micro, ni le retour sur les données. La visualisation ouvre des pistes d'analyse nouvelles que le chercheur ne percevait pas avec autant d'évidence lors de l'étude des données à l'écoute ou par lecture des transcriptions.

[80] Bourdieu, P., Passeron, J.-C., & Chamboredon, J.-C. (1968). *Le métier de sociologue*.
[81] Accardo, A., Balazs, G., & Beaud, S. (1993). *La misère du monde*.

3.4 Méthodologie de collecte

Articles et livre :	
	○ 2006, *Corpus oraux, Guide des bonnes pratiques*, https://halshs.archives-ouvertes.fr/halshs-00355472
Communications orales :	
	○ 2009, Les enquêtes sociolinguistiques à Orléans (1970-2009), l'entretien en questions, https://halshs.archives-ouvertes.fr/halshs-01165950 ○ 2006, Constitution et exploitation d'un grand corpus de "données situées" Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), https://halshs.archives-ouvertes.fr/halshs-01165954
Documents :	
	○ Les techniques de prises de son, site ESLO : http://eslo.huma-num.fr/index.php/pagemethodologie?id=70 ○ Guide du matériel ESLO2 : http://www.nakala.fr/data/11280/e177902b ○ Plaquette de présentation ESLO2 : http://www.nakala.fr/data/11280/6d42a893 ○ Affiche ESLO2 : http://www.nakala.fr/data/11280/81b2b601 ○ Formulaire de consentement ESLO2 : http://www.nakala.fr/data/11280/df5c9365

3.4.1 L'entretien en question *[aspects juridiques]* *[retour]*

Nous l'avons précisé, le projet ESLO2 s'appuie sur une analyse de l'évolution des cadres théoriques depuis ESLO1. Ceux-ci ont un effet important sur l'architecture du corpus mais aussi sur la méthodologie de collecte des différents modules et des entretiens en premier lieu. Ce travail, là aussi fortement empreint de réflexivité, a été développé au fil de différentes communications sur cette approche (Baude et Perrot 2009[82]).

Les principaux cadres théoriques à l'origine de l'évolution méthodologique sont clairement identifiables :

- **La description par W. Labov du paradoxe de l'observateur** [Labov 73, 82] va orienter les recherches sociolinguistiques vers la méthodologie de l'observation participante [Bouziri 2000].
- **Les recherches en analyse de la conversation** en linguistique et en ethnométhodologie vont se construire autour de l'opposition frontale entre données provoquées et non provoquées par le chercheur.
- **Les études conduites en sociologie et en anthropologie sur les techniques d'enquête** [Passeron et Bourdieu 68, Bourdieu 93, Beaud 96] conjointes aux travaux sur la légitimité et l'insécurité linguistique [Bourdieu 82, Encrevé et de Fornel 83] ont conforté les réticences envers cette technique d'enquête, notamment pour ce qui a trait à la naturalité des productions linguistiques (polémique sur la transcription Bourdieu 93, Lahire-Beaud 96].

[82] https://halshs.archives-ouvertes.fr/halshs-01165950

- **Les travaux sur la linguistique des genres** [Biber 98] et sur une typologie des productions liée à une typologie des situations de communications [Koch Peter & Wulf Oesterreicher, 2001] confinent les entretiens à un contexte de production linguistique particulier.

- **Les évolutions technologiques** (qualité et discrétion du dispositif d'enregistrement, traitement des données numériques et même développement de la vidéo favorisant les travaux en linguistique interactionnelle [Goodwin 81, Mondada 2001-06]).

A ceci s'ajoute une évaluation des résultats obtenus dans ESLO1 et notamment l'échec de la représentativité du panel par manque de réponses de certaines catégories sociales de locuteurs.

> « *Ce sont les services de l'INSEE qui sur les instructions des membres de l'équipe d'enquête ont procédé au tirage au sort de six cents témoins répartis également entre six catégories socioprofessionnelles, le pourcentage d'échecs prévus au départ étant de 50%. En réalité seulement 147 témoins appartenant à cet échantillon ont été interviewés (…)* ».

Il n'y a rien d'étonnant dans cet échec, *« L'enquête ethnographique nous apprend très rapidement que toute personne sociale n'est pas "interviewable", qu'il y a des conditions sociales à la prise de parole ». [S. Beaud 1997:234]*[83]

Il s'agit néanmoins de définir le genre de l'entretien qui dans ESLO2 correspond à des objectifs précis. L'entretien est un genre particulier. *Le registre du langage utilisé par les témoins est probablement le mieux caractérisé comme un langage semi-formel, dans le sens de la notion labovienne de careful speech* (voir Labov 1972). [De Jong 1994:]. Il offre l'opportunité d'une stabilité dans la méthodologie de l'enquête nécessaire à la comparabilité au sein d'échantillons. Et il permet par *réentrée* des informations recueillies un affinement de l'échantillonnage à l'aide de critères multiples.

Contrainte technique et formatage dû à la méthodologie

En premier lieu, le dispositif technique d'enregistrement, qui devient un élément essentiel de la qualité d'un corpus oral, constitue ici une opportunité. En effet, un corpus sociolinguistique est en premier lieu une "captation" de productions linguistiques qui requièrent un minimum de compétences techniques. Les enregistrements ESLO2 ont été réalisés à l'aide d'enregistreurs numériques de marque Marantz, modèle PMD661-MK2. Ce modèle a pour entrées un ou deux micros externes et un micro interne. Le format

[83] BEAUD, S., & WEBER, F. (1997). *Guide de l'enquête de terrain: produire et analyser des données ethnographiques*.

d'enregistrement des fichiers est celui standard du wave, 16 bits, 44100Hz, stéréro. Les enregistrements sont réalisés en stéréo, ce qui facilite la transcription (possibilité d'isoler l'un des deux canaux). Pour les entretiens nous avons utilisé des micros externes (micro-cravate) de marque AKG, modèle C417[84]. Un guide complet sur le matériel technique est disponible (http://www.nakala.fr/data/11280/e177902b).

Si pour ESLO1 le "micro autour du cou" ou le "micro caché" aboutissaient à une restitution de mauvaise qualité, pour ESLO2, le micro-cravate et la captation individuelle/double est d'une qualité acoustique très supérieure. Paradoxalement, l'entretien est l'une des rares situations ou le micro est accepté comme un élément "naturel" (vs micro en situation non provoquée par le chercheur).

La constitution d'un corpus représentant une masse de données nécessite un formatage important afin de permettre l'usage d'outils d'extraction et de comparaison des informations. L'entretien est une technique qui permet ce genre de travail :

> *« Un autre inconvénient de la situation d'interview est son caractère artificiel. Pour pallier ce défaut, il faut faire appel à d'autres situations de discours plus naturelles, où les emplois de la langue sont plus spontanés : discussions de groupes, conversations de famille ou dans la rue, etc. Mais ce que ces situations gagnent en spontanéité, elles le perdent tant sur le plan de la rigueur méthodologique (contrôle des variables situationnelles) et celui de la qualité technique de l'enregistrement, sans parler de la difficulté à réunir les conditions d'une authentique spontanéité. »* [Blanc & Biggs 1971: 17]

Cependant on ne peut contester les effets de « tamisage » des données produites sur questionnaire [Latour 1993, Mondada 1998:49]. Celui-ci permet une domestication du terrain rendu conforme aux ordres et phénomènes recherchés et aux analyses prévues. Il garantit :

- l'épuration des données (élimination des bruits),
- la compatibilité des données avec les analyses, les calculs, la formalisation,
- la comparabilité de données recueillies dans des lieux et à des moments différents.

Dans ESLO2 l'entretien est considéré comme une « ressource » mais aussi comme un « évènement interactionnel » :

- L'entretien ressource

Si l'entretien permet, tout en recueillant des productions linguistiques, de favoriser le traitement quantitatif des données en filtrant, unifiant, formalisant le terrain à tous les niveaux :

- classification du locuteur,

[84] Cf fiche technique sur le site ESLO : http://eslo.huma-num.fr/index.php/pagemethodologie?id=70

- description de la situation,
- découpage du contenu (110 questions repérées dans ESLO1 en 3 thèmes),
- découpage simplifié des tours de paroles avec un respect affirmé des conventions d'alternance,

il est aussi et avant tout un évènement à part entière dont l'analyse nécessite un cadre théorique et une méthodologie relevant de la sociolinguistique.
Pour ces deux raisons la structure du corpus des Eslos(1&2) sous forme numérique est conçue comme un tout contenant l'ensemble des données : sons, transcriptions, métadonnées décrivant le contexte, métadonnées de catalogage et d'indexation (base de données XML native).

L'entretien évènement:

- L'entretien est un micromarché linguistique ou s'exercent des stratégies de défense et de construction d'un capital linguistique [Bourdieu 1982]. Les informations fournies lors d'un entretien doivent être interprétées à la lumière de ces stratégies.
- Cette approche recèle une critique de l'implicite quantitatif du travail par entretien et le plaidoyer pour "l'entretien ethnographique" [Passeron, Beaud, Weber,…], [Bourdieu 1993]
- Il y a un refus d'objectivation (totale) des informations recueillies. Ce sont des interprétations co-produites conjointement par l'enquêteur et l'informateur au sein d'un évènement communicationnel. L'entretien configure le contexte et la référence qu'il produit. « *Ainsi contrairement à un entretien lors d'une conversation on répond rarement de façon définitive à une question, mais on élabore progressivement, au cours de formulations, de négociations, la réponse qui peut se transformer au fil du temps* » [Mondada 1998:59]

"*Ce que les acteurs sociaux se disent n'est pas seulement une ressource privilégiée du sociologue pour savoir ce qui se passe, se pense, se sent, etc. mais une partie intrinsèque de ce qui se passe*" (Widmer 1985:65).

Mode d'approche et catégorisation des rôles :

« Le portrait sonore de la ville » a été utilisé dès ESLO1 afin de donner un objectif facilement identifiable par les témoins :

« La manière d'aborder les témoins était également assez délicate : les buts de l'enquête devaient sembler plausibles sans éveiller la méfiance ni surtout la suspicion que l'investigation pouvait porter sur la façon de parler des sujets, ce qui aurait faussé complètement l'enquête. La solution adoptée a consisté à donner à l'enquête un tour journalistique. L'équipe se présentait comme un groupe d'universitaires britanniques et français désireux de réaliser le "Portrait sonore d'Orléans" ». [Blanc & Biggs 1971:21]

ESLO1:123

RV 252*: entrez par là remarquez je suis pas du tout au courant de de ce qui s'agit hein j'ai j'ai reçu des des papiers des lettres euh disant que vous cherchiez à à interviewer ou je sais pas*

AR: *non euh c'est l'équipe franco-britannique qui cherche à faire un portrait euh ça n*

Pour ESLO2 cette démarche a été renforcée et un flyer et des affiches ont été spécialement conçues pour lancer de la sorte la campagne de collecte :

Cette démarche a donné lieu à une reconsidération de l'usage du questionnaire au profit d'un entretien très peu directif et contenant assez peu de questions afin de diminuer la violence symbolique et souvent linguistique de la situation d'entretien :

> *« On a souvent examiné le couple Q-R, mais peut-on mener à terme l'analyse pragmatique de cet enchainement discursif sans y repérer l'inscription linguistique d'un rapport social exigeant que l'on remonte de la question à l'acte de questionnement ? (…) la question bien qu'elle se présente comme une demande d'information est une prise effectuée sur un autre sujet (…) »* [Encrevé 1983:6]

En accord avec les avancées de la sociologie :

> *« Plusieurs dizaines d'années d'exercice de l'enquête sous toutes ses formes, de l'ethnologie à la sociologie, du questionnaire dit fermé à l'entretien le plus ouvert, m'ont ainsi convaincu que cette pratique ne trouve son expression adéquate ni dans les prescriptions d'une méthodologie souvent plus scientiste que scientifique, ni dans les mises en garde scientifiques des mystiques de la fusion affective. C'est pourquoi il me parait indispensable d'essayer d'expliciter les procédures que nous avons mises en œuvre dans la recherche dont nous livrons ici les résultats.*
>
> *On a donc essayé d'instaurer une relation d'écoute active et méthodique, aussi éloignée du pur laisser-faire de l'entretien non directif que du dirigisme du questionnaire »* [Bourdieu 1993:903].

C'est ainsi que le questionnaire contenant plus d'une centaine de questions, utilisé dans ESLO1, a été abandonné au profit d'une trame pour une « menée » non directive des entretiens. Cela donne des passages dont le chercheur ne sort pas toujours à son avantage :

ESLO2DD4 : hein par exemple j'ai fait un un bourguignon euh j'ai fait un bourguignon quand donc on est on est quel jour on est lun- lun- lundi euh j'ai fait un bourguignon samedi j'ai invité Véronique et Jean qui sont venus en manger hier soir il m'en reste vous voulez venir en manger ?
OB [rire] non mais z- z- moi je suis a- attendu par euh euh p- p-
OB + ESLO2DD4 1 : ap- après
2: non mais la famille
OB : après b- euh la famille euh
ESLO2DD4 : et il reste du bourguignon suffisamment pour
OB : ça ça c'est gentil mais euh
ESLO2DD4 + OB : 1: vous voulez pas de mon bourguignon ?
2: on va faire ça
OB : ah si il a l'air très bon votre bourguignon ça

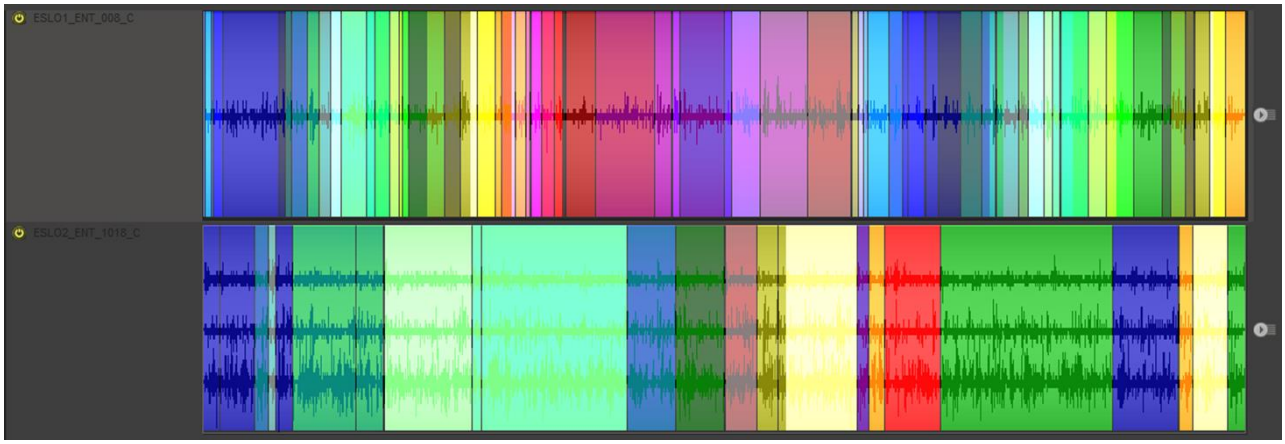
La « menée » d'entretien[85] d'ESLO2 est construite autour d'un objectif clairement identifiable de portrait sonre de la ville, un thème cohérent, repérable et justifiant le recueil "d'avis". C'est également une unité d'analyse historique, géographique, sociologique repérable et maitrisable. Elle permet le recentrage sur l'individu : l'agent dans sa cité, le portrait de la ville se dessinant à travers le portrait des gens dans la ville, leurs expériences urbaines quotidiennes). Le propos est centré sur les pratiques du témoin sur lesquelles l'enquêteur fait retour sans s'interdire de relancer le témoin sur des thématiques évaluatives. L'objectif est de partir des pratiques concrètes (incorporées, contextualisées, liées à un déplacement dans l'espace, à des conduites cognitives etc.) afin de dresser un

[85] https://www.nakala.fr/data/11280/6381d925

portrait sociologique fin. Une thématique urbaine a pour avantage de fournir des indicateurs nombreux et peu surveillés dont la structure est facilement identifiable.

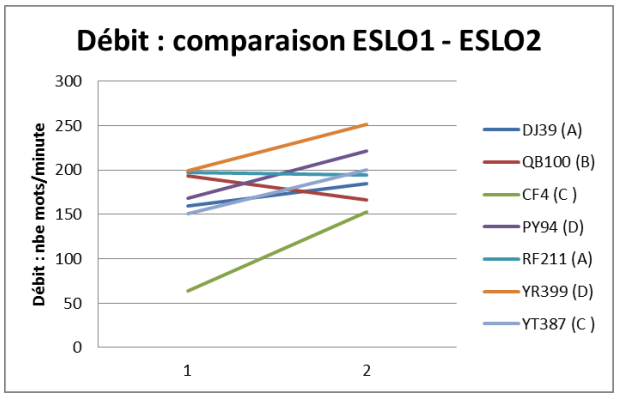
3.4.2. Premiers aperçus d'une analyse diachronique [retour]

La présence dans le corpus ESLO2 d'un module « diachronique » constitué de sept locuteurs ESLO1 réenregistrés à quarante années de distance par Annie Chesneau dans le cadre d'un doctorat, permet de tester quelques résultats des modifications introduites dans la méthodologie. Le graphique suivant, obtenu à partir du logiciel de traitement d'enquête SONAL, permet de visualiser la fluidité de l'entretien d'ESLO2 comparé à l'entretien ESLO1. Les couleurs correspondent aux thématiques structurées par le questionnaire :



La différence est significative et montre un entretien ESLO2 plus fluide et moins saccadé que l'entretien ESLO1. Nous verrons par la suite l'incidence de cette méthodologie de collecte sur les analyses. Toutefois on peut d'ores et déjà corréler ces données avec celles concernant la comparaison du débit de parole entre ESLO1 et ESLO2 à partir des entretiens des locuteurs présents dans les deux corpus.

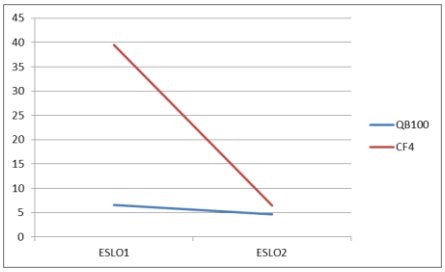
ESLO1	Nbre mots	durée	durée	débit	Locuteur	AM	ESLO2	nbre de mots	durée	durée	débit
ESLO1_ENT_003_C	9326	00:58:32	58,53	**159,33**	DJ39	A	ESLO2_DIA_1221_C	9861	00:53:20	53,33	**184,89**
ESLO1_ENT_017_C	15252	01:19:00	79	**193,06**	QB100	B	ESLO2_DIA_1222_C	8761	00:52:38	52,63	**166,45**
ESLO1_ENT_048_C	6514	01:41:51	101,85	**63,96**	CF4	C	ESLO2_DIA_1223_C	12137	01:19:33	79,55	**152,57**
ESLO1_ENT_115_C	10907	01:04:46	64,77	**168,4**	PY94	D	ESLO2_DIA_1224_C	8637	00:39:02	39,03	**221,27**
ESLO1_ENT_121_C	24687	02:05:29	125,48	**196,74**	RF211	A	ESLO2_DIA_1225_C	9120	00:46:57	46,95	**194,25**
ESLO1_ENT_149_C	21353	01:47:17	107,28	**199,03**	YR399	D	ESLO2_DIA_1226_C	22211	01:28:32	88,53	**250,88**
ESLO1_ENT_150_C	6495	00:43:11	43,18	**150,41**	YT387	C	ESLO2_DIA_1227_C	12853	01:04:23	64,38	**199,63**



Les chiffres montrent une évolution significative du débit entre ESLO1 et ESLO2 pour la plupart des locuteurs. On peut en déduire une évolution qui correspond aux études actuelles sur le débit en français. Toutefois deux données supplémentaires permettent d'affiner cette analyse : d'une part, le débit augmente moins chez les locuteurs appartenant aux catégories sociales les plus élevés ; d'autre part, les chiffres s'avèrent très différents si on retire les pauses et les silences. On peut donc avancer l'hypothèse que ce qui caractérise la différence entre ESLO1 et ESLO2 c'est bien la diminution des temps de silence présents dans les entretiens, éléments à rapprocher de la recherche d'une menée fluide de la discussion en opposition avec un questionnaire dispensé avec une certaine rigidité.

Ces données sont d'ailleurs confirmées par une analyse du temps de silence sur quatre entretiens. Elle est considérable pour l'un des locuteurs et reste significative pour le second :

	ESLO1	ESLO2
QB100	6,58	4,71
F4	39,54	6,42



Cette brève analyse sur fond d'approche réflexive de la méthodologie a permis de conforter les choix de collecte décidés pour ESLO2.

3.4.3. Aspects juridiques : le recueil de consentement éclairé

[Guide des Bonnes Pratiques] [retour]

Les aspects juridiques complétés par une démarche éthique rendue possible par une approche réflexive ont été abordés parallèlement aux travaux en cours depuis une quinzaine d'années dans le domaine des corpus oraux en particulier et des archives scientifiques en général. Les grands principes qui guident les solutions appliquées dans le projet ESLO proviennent du *Guide des bonnes pratiques 2006* tels qu'ils sont présentés dans le chapitre consacré à cette question. Nous donnerons dès à présent un exemple concret ayant une influence directe sur la méthodologie de collecte et sur le traitement du corpus.

L'apport principal du *Guide* est de permettre un repérage des cadres juridiques, pour lesquels l'expertise d'Isabelle de Lamberterie a été décisive, tout en les confrontant à une explicitation de la démarche du chercheur.

Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012 (Eshkol et al 2011 :21)

À l'époque de la conservation pérenne et de la diffusion des archives numériques il est important de prendre en compte l'ensemble des aspects juridiques dès la conception du projet. Nous avons bénéficié des réflexions et recommandations émanant du groupe de travail du ministère de la Culture et du CNRS16 : Corpus oraux. Guide des bonnes pratiques 2006*. Les problèmes rencontrés se concentrent sur deux grands domaines juridiques : le respect de la vie privée et la protection de la propriété intellectuelle.*

Les deux questions principales gouvernées par des aspects juridiques sont les suivantes : le corpus contient-il des données personnelles ? Quels éléments relèvent du droit d'auteur ?

La première question a un impact direct puisqu'une réponse affirmative implique le recueil du consentement de la personne impliquée.

Corpus oraux, Guide des bonnes pratiques 2006 (2006 :41)

La création d'un corpus passe le plus souvent par la collecte de données. Celles-ci pouvant être des données personnelles, cette collecte doit être faite dans le respect de la loi Informatique et libertés : licéité et loyauté, information préalable, obtention du consentement des personnes concernées (voir fiche Consentement), respect des finalités annoncées …

La seconde question sera traitée plus précisément dans le chapitre sur les licences. La position adoptée par l'ensemble de l'équipe du projet a été de considérer que le corpus est produit dans le cadre d'un projet du laboratoire et que s'il convient de mentionner les droits

moraux de chaque participant, les droits patrimoniaux ne s'appliquent qu'à l'entité laboratoire. Cette position a été étendue aux locuteurs participants, ce qui sur ce point nécessitait de recueillir leur autorisation pour une utilisation dans le cadre des données de la recherche.

Corpus oraux, Guide des bonnes pratiques 2006 (2006 :41)

Les droits patrimoniaux se résument en un droit exclusif au profit de l'auteur (ou des titulaires) ou des ayants droit (bénéficiaire d'une cession, héritiers...) d'autoriser ou interdire la reproduction ou la communication au public de l'œuvre protégée. Si le corpus oral est une œuvre, toute reproduction (la numérisation est pour le droit une reproduction) et toute mise à disposition du public (sur un site Internet comme sur tout autre support) nécessitent l'autorisation expresse de l'auteur ou du titulaire de droit.

*Quant aux prérogatives du droit moral, toujours attachées à la personne physique créatrice de l'œuvre protégée, elles sont au nombre de quatre : le *droit de divulgation, le *droit de repentir et de retrait, le *droit à la paternité et le *droit au respect de l'œuvre. Chacun de ces droits est applicable aux corpus oraux.*

Pour que les droits patrimoniaux et moraux s'appliquent, il faut néanmoins qu'il y ait un ou plusieurs auteurs d'une « œuvre » et donc qu'un corpus soit défini par une forme spécifique (originale) et relevant d'une activité créatrice.

Corpus oraux, Guide des bonnes pratiques 2006 (2006:39)

Quelles sont les conditions pour qu'un corpus soit protégé ? Il y en a trois.

Il faut en premier lieu qu'il corresponde à l'exigence d'une activité créatrice : un travail de compilation d'informations n'est pas protégé en soi.

Pour être protégé, il est par ailleurs indispensable que le corpus ait une forme définie. Ce qui est protégé, ce n'est pas le contenu du corpus mais son enveloppe, son architecture.

Enfin, la forme du corpus doit répondre à la condition d'être originale.

Ces éléments juridiques se traduisent, dans le cadre du projet ESLO, par le recueil du consentement éclairé des personnes enregistrées. Nous le verrons par la suite, cette disposition a été complétée par une procédure d'anonymisation partielle et par des licences de distribution du corpus. Il n'en reste pas moins que la première des opérations, celle si

souvent oubliée ou peu maitrisée dans les projets d'enquête sociolinguistique, reste le recueil du consentement éclairé.

Il s'agit d'un point particulièrement délicat dans le cadre d'enquêtes sociolinguistiques où le chercheur souhaite collecter des paroles hors d'un contexte d'autosurveillance. Il y a alors un paradoxe accru à vouloir informer avec précision le locuteur des objectifs d'analyse linguistique tout en espérant qu'il formule des énoncés dans un cadre informel. Le sociolinguistique sait mieux que quiconque que la nature sociale même de la langue fait que tout contexte est celui d'un marché linguistique au sein duquel les locuteurs ont conscience que leurs productions sont évaluées selon une loi de formation des prix fortement gouvernée par la structure de la société.

La solution adoptée dans le cadre du projet ESLO2 se divise en trois parties :

1° Construire un projet participatif

Le premier élément de solution est d'élaborer un projet qui dépasse l'enquête linguistique. C'est ainsi que le projet ESLO est intégré à un projet plus vaste qui consiste à dresser le portrait sonore de la ville d'Orléans. Il est plus simple d'informer les personnes enregistrées des objectifs du projet sans attirer l'attention sur l'aspect linguistique. Cette idée présente dès ESLO1 a été accusée dans ESLO2 et elle est à l'origine de différents projets réalisés dans le cadre d'actions culturelles et artistiques. C'est ainsi que de nombreux enregistrements d'ESLO2 ont été réalisés dans le cadre de projets menés par des artistes selon une finalité bien identifiée par les locuteurs. Dans cette perspective, le choix du dispositif technique (micro-cravate) est adapté à la production d'évènements culturels et artistiques (médiés) et identifiés comme tels par les participants.

2° L'information préalable à l'enregistrement

Le deuxième élément de solution a consisté à donner une information sommaire, préalable à l'enregistrement. L'objectif est d'informer le locuteur de la captation de ses propos et du moment à partir duquel celle-ci commence. Le locuteur est ainsi averti sans qu'on puisse considérer qu'il s'agit d'un consentement éclairé.

3° Le consentement éclairé post-enregistrement

Le troisième élément de solution est d'informer précisément le locuteur après l'enregistrement afin de ne pas exercer d'effet de ces informations sur le déroulé de l'interaction. Le consentement peut alors être détaillé, tant sur les conditions de diffusion que d'exploitation des enregistrements. Le formulaire écrit énumère les objectifs et les opérations que subiront les données enregistrées, y compris leurs conditions de diffusion :

En conséquence, j'autorise :

	Oui	Non
1° L'enregistrement audio de l'entretien réalisé le……..	❑	❑
2° L'utilisation de cet enregistrement sous sa forme sonore ainsi que sous ses formes transcrites pour :	❑	❑

- la recherche scientifique (travaux d'analyse divers, thèses, articles scientifiques, communications lors de colloques,…),
- des usages d'enseignements (utilisation lors de cours, manuels et autres matériels pédagogiques),
- la recherche sur les technologies de la langue (reconnaissance de la parole, synthèse vocale, …)
- une diffusion à d'autres équipes de recherche de la communauté scientifique,
- une diffusion "grand public" (ouvrage de vulgarisation, site internet d'archive…).

Par la présente j'autorise l'utilisation scientifique et non commerciale de l'enregistrement et de sa transcription par le Laboratoire Ligérien de Linguistique.

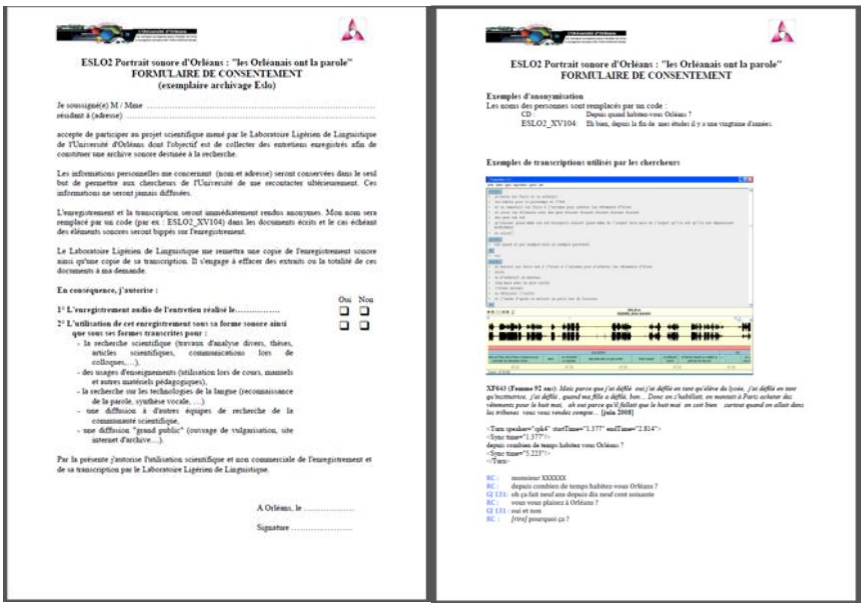
Le formulaire contient également des informations circonstanciées sur les conditions d'anonymisation et sur la forme que revêt une transcription réalisée par le laboratoire :

> Les informations personnelles me concernant (nom et adresse) seront conservées dans le seul but de permettre aux chercheurs de l'Université de me recontacter ultérieurement. Ces informations ne seront jamais diffusées.
>
> L'enregistrement et la transcription seront immédiatement rendus anonymes. Mon nom sera remplacé par un code (par ex : ESLO2_XV104) dans les documents écrits et le cas échéant des éléments sonores seront bippés sur l'enregistrement.

Le document est réalisé en deux exemplaires ce qui permet au locuteur de conserver une trace de son engagement et l'autorise à demander la destruction de tout ou partie de l'enregistrement réalisé :

> Le Laboratoire Ligérien de Linguistique me remettra une copie de l'enregistrement sonore ainsi qu'une copie de sa transcription. Il s'engage à effacer des extraits ou la totalité de ces documents à ma demande.

Document intégral (http://www.nakala.fr/data/11280/df5c9365) :



ESLO2 Portrait sonore d'Orléans : "les Orléanais ont la parole"
FORMULAIRE DE CONSENTEMENT
(exemplaire archivage Eslo)

Je soussigné(e) M / Mme .....
résidant à (adresse) .....

accepte de participer au projet scientifique mené par le Laboratoire Ligérien de Linguistique de l'Université d'Orléans dont l'objectif est de collecter des entretiens enregistrés afin de constituer une archive sonore destinée à la recherche.

Les informations personnelles me concernant (nom et adresse) seront conservées dans le seul but de permettre aux chercheurs de l'Université de me recontacter ultérieurement. Ces informations ne seront jamais diffusées.

L'enregistrement et la transcription seront immédiatement rendus anonymes. Mon nom sera remplacé par un code (par ex : ESLO2_XV104) dans les documents écrits et le cas échéant des éléments sonores seront floppés sur l'enregistrement.

Le Laboratoire Ligérien de Linguistique me remettra une copie de l'enregistrement sonore ainsi qu'une copie de sa transcription. Il s'engage à effacer des extraits ou la totalité de ces documents à ma demande.

En conséquence, j'autorise :

	Oui	Non
1° L'enregistrement audio de l'entretien réalisé le.....	☐	☐
2° L'utilisation de cet enregistrement sous sa forme sonore ainsi que sous ses formes transcrites pour :	☐	☐

- la recherche scientifique (travaux d'analyse divers, thèses, articles scientifiques, communications lors de colloques,...),
- des usages d'enseignements (utilisation lors de cours, manuels et autres matériels pédagogiques),
- la recherche sur les technologies de la langue (reconnaissance de la parole, synthèse vocale,...)
- une diffusion à d'autres équipes de recherche de la communauté scientifique,
- une diffusion "grand public" (ouvrage de vulgarisation, site internet d'archive...).

Par la présente j'autorise l'utilisation scientifique et non commerciale de l'enregistrement et de sa transcription par le Laboratoire Ligérien de Linguistique.

A Orléans, le .....

Signature .....

ESLO2 Portrait sonore d'Orléans : "les Orléanais ont la parole"
FORMULAIRE DE CONSENTEMENT

Exemples d'anonymisation
Les noms des personnes sont remplacés par un code :
CD : Depuis quand habitez-vous Orléans ?
ESLO2_XV104 : Eh bien, depuis la fin de mes études il y a une vingtaine d'années.

Exemples de transcriptions utilisés par les chercheurs

La prise en charge de ces aspects juridiques ne peut se réduire à un respect des textes législatifs. Cela nécessite tout d'abord une approche éthique et une explicitation de la démarche du chercheur qui relèvent toutes deux d'une démarche réflexive. Ensuite, c'est la méthodologie de la collecte qui est décidée par cette approche. Il convient de construire un projet aux objectifs clairement identifiables et d'être en mesure de les rendre disponibles à tous les participants, puis d'organiser le recueil de consentement en parfaite adéquation avec un protocole détaillé. Enfin la structuration même des données, leur conservation et leur diffusion sont concernées par ces choix initiaux. Là encore, la maitrise de la chaine de traitement d'un corpus oral est cruciale tout comme il est essentiel qu'elle soit prise en compte dans l'effet qu'elle a sur la production des données et leurs analyses.

3.5 Transcrire les ESLOs : codage et annotation première [retour]

Articles :	
	○ 2006, Constituer et exploiter un grand corpus oral, choix et enjeux théoriques, le cas des ESLO, https://halshs.archives-ouvertes.fr/halshs-01162506 ○ 2011, Un grand corpus oral disponible, le corpus d'Orléans 1968-2012, https://halshs.archives-ouvertes.fr/halshs-01163053v1
Communications orales :	
	○ 2005, Transcrire les bonnes pratiques des linguistes https://halshs.archives-ouvertes.fr/halshs-01162548v1 ○ 2008, de la variation à la norme, les effets de codage dans les ESLOs https://halshs.archives-ouvertes.fr/halshs-01165953v1 ○ 2009, Sociolinguistique et transcription, https://halshs.archives-ouvertes.fr/halshs-01165951v1 ○ 2011, Transcrire, la norme, la variation, le linguiste, https://halshs.archives-ouvertes.fr/halshs-01165948v1
Documents :	
	○ Guide du transcripteur V1 : http://www.nakala.fr/data/11280/caea0537 ○ V2 : http://www.nakala.fr/data/11280/7445014a ○ V3 : http://www.nakala.fr/data/11280/96887cc0 ○ V4 : http://www.nakala.fr/data/11280/2cf7a33a ○ Lexique ESLO V-2014 : http://www.nakala.fr/data/11280/55834601 ○ Page site ESLO, Transcription: http://eslo.huma-num.fr/index.php/pagemethodologie?id=71 ○ http://www.tei-c.org/index.xml

3.5.1 De l'oral à l'écrit ? [retour]

Plus de trente ans après les travaux de Claire Blanche-Benveniste[86] et ceux de Halliday[87] et Ochs[88] notamment, la question de la transcription reste un enjeu majeur et une source de difficultés pour toute étude sur l'oral.

La difficulté majeure reste celle du codage à utiliser qui ne peut se ramener à la notation conventionnelle de la langue écrite :

> « *On ne peut pas étudier l'oral par l'oral, en se fiant à la mémoire qu'on en garde. On ne peut pas sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer des morceaux. (…) On rencontre alors une série de difficultés, du fait que l'écriture n'est pas le simple instrument de transposition de l'oral qu'une approche naïve voudrait y voir.*» (Blanche-Benveniste, 1997:24-25[89])

[86] Blanche-Benveniste, C., & Jeanjean, C. (1987). *Le français parlé: transcription et édition*.
[87] Halliday, M. A. K. (1985). *Spoken and written language*.
[88] Ochs, E., & Schieffelin, B. B. (Éd.). (1979). *Developmental pragmatics*.
[89] Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*.

A cette difficulté s'en ajoute une deuxième quand il s'agit d'un grand corpus oral, impliquant une masse de données volumineuse.

Enfin, l'enjeu de l'interopérabilité des données apporte une troisième difficulté : la transcription produit-elle des données secondaires résultant d'une forme d'analyse ou doit-elle être considérée comme une donnée primaire, source de l'étude ?

L'objectif méthodologique de comparer des segments sonores à partir d'un codage normalisé a été clairement défini par la linguistique de corpus informée des recherches conduites en TAL :

> "*Il est peu probable qu'une science se trouve en position de développer des théories profondes pour expliquer ses données avant qu'il existe un cadre faisant consensus (agreed scheme) sur la manière d'identifier et de noter ces données*". (Sampson, 2001[90])

> « *Pour qu'une base textuelle permette l'extraction à la demande des « documents » en fonction d'une utilisation donnée, il importe que chacune des unités élémentaires qui la constituent soit « autonomisable ». On doit posséder suffisamment d'informations fines sur elle pour pouvoir l'extraire de la base et l'assembler avec d'autres éléments de la même base ou d'autres bases sans perdre ces renseignements qui sont indispensables pour interpréter les contrastes et les convergences manifestés dans le corpus qui vient d'être rassemblé. Ces renseignements doivent couvrir à la fois la description précise du contexte de production du composant et une caractérisation en termes de domaine (thématique) et de « genre » (au sens indiqué supra).* » (Habert, 2001)

Si cet objectif est largement partagé, il a été fortement relativisé par les linguistes plus proches d'une linguistique de corpus descriptiviste et de terrain :

> *Les recherches en matière de normalisation et de planification des langues historiques révèlent essentiellement deux composantes du processus de standardisation : la sélection et la codification (…). Cette normalisation orthographique vise beaucoup moins à optimiser les correspondances phonèmes : graphèmes qu'à unifier des modèles et des traditions graphiques coexistants ou à en éliminer d'autres. (…) Les lignes de partage tracées par des écritures, voire des systèmes d'écriture différents, sont encore plus profondes, puisque ce sont des facteurs socioculturels,*

[90] SAMPSON, G. (2001). *Empirical linguistics*.

> *notamment religieux ou politiques, qui déterminent le choix de l'écriture: que l'on pense p. ex. au serbo-croate ou au moldave-roumain, où s'opposent, respectivement, les écritures latine et cyrillique, et à des cas extrêmes comme le hindi-ourdou réalisé soit en écriture devanagari, soit en écriture perso-arabe (cf. Comrie 1987, 473-476).* (Koch & Oesterreiche, 2001[91])

> *« Pour le linguiste praticien averti de la transcription, le rêve du copiste qui saisirait la matière vivante telle qu'elle est, objectivement, et la restituerait ipso facto, est illusoire. (...) Le réel transcrit n'est jamais le réel émis mais un réel perçu et traduit. Au cours de ce processus, on assiste à une série de filtrages, d'appauvrissements, d'interprétations. »* [Giovannoni et Savelli 1990]

Il s'agit bien, comme le soulignait Ochs dès 1979, de considérer la transcription comme une étape d'analyse porteuse de cadre théorique :

> *« Notre argument consiste à dire que la transcription est un moment particulier de la recherche où des catégorisations exogènes peuvent être produites par le transcripteur sans être nécessairement maîtrisées, avec des effets structurants sur l'analyse qui va suivre. Ce risque nous porte à souligner en retour l'importance de la pratique de la transcription comme moment central de l'analyse – loin de la réduire à un « simple » problème technique et à une pratique professionnelle subordonnée. (*Mondada, 2002[92])

> « *La transcription n'est donc pas un "mobile immuable" (Latour, 1986) qui une fois établi pourrait à la fois circuler dans des réseaux socio-techniques et préserver ses propriétés essentielles. En tant que tel, elle pourrait, grâce à sa stabilisation d'une version reconnue et fiable des faits au-delà des controverses et des désaccords, imposer une forme de facticité et de réalisme de la "donnée" (Latour , 1989). Mais elle n'y parvient pas (…).*
> *Dans ce sens la transcription est plutôt un "objet intermédiaire" (Vink, 1999), un "objet frontière", un "boundary object" (Star & Griesmer, 1989) c'est-à-dire un objet qui traverse des communautés, qui relie des chercheurs et qui est partagé entre eux, mais auquel différents sens sont attribués, pris dans des pratiques de production et de lecture qui, malgré les appels de la standardisation, demeurent hétérogènes. Son statut de "boundary object"*

[91] Koch P. & Oesterreicher W., 2001, « Langage parlé et langage écrit », Lexikon der romanistischen Linguistik, tome 1, Max Niemeyer Verlag, Tübingen
[92] Mondada, Transcription practices and categorization effects, 2002

> *éclaire les controverses qui ont lieu à son propos dans le cadre des tentatives de standardiser (voir Pickering 1992). (*Mondada 2008 :81)[93].

Cette position peut aller jusqu'à une critique qui, si elle est juste, n'en pose pas moins une contrainte difficile à résoudre si on intègre un objectif d'interopérabilité :

> *« Toutes les modalités de transcription ont, certes, des qualités, mais tout autant des limites, et elles sont toutes d'une façon ou d'une autre idéologisées, la standardisation radicale comme les autres. C'est que les choix sont davantage motivés par des objectifs analytiques que par une quelconque volonté de vérité ou de fidélité, derrière lesquelles on s'abrite en général. Il faut donc mesurer les implications d'une transcription, et être conscient du contexte socio-politique de sa réception, car la transcription constitue toujours l'exercice d'un pouvoir : "conscience et responsabilité", plutôt qu'illusion "de la neutralité scientifique" selon les mots de Buchholtz 2000 p.1461". »* (Gadet 2008 :44)[94]

> « *Notre conclusion essentielle est qu'une transcription est toujours effectuée en vue d'un projet de recherche : elle ne serait ainsi pas réutilisable sans reformulation par d'autres chercheurs avec d'autres objectifs et il faudrait la reconsidérer en tenant compte de ses propres objectifs de recherche, de façon à s'approprier la démarche autant que son produit le transcript. Il n'y a d'ailleurs là rien de surprenant car si la transcription est bien un geste théorique alors il prend place dans une chaine de gestes théoriques (…). Pas plus qu'il n'y a d'analyse tous azimuts et tous objectifs, il ne saurait y avoir de transcription tous azimuts et tous objectifs, aussi minimale puisse-t-elle paraitre ; et il faut renoncer aux rêves d'harmonisation.* » (Gadet 2008 :46)[95]

3.5.2 Annotation de niveau zéro [retour]

Face à ces difficultés, la solution adoptée pour la transcription des corpus ESLOs sera de deux niveaux.

L'impact des aspects théoriques induits par la transcription est indéniable et nous partageons l'avis de ceux qui pensent qu'une analyse de données orales ne peut se satisfaire

[93] Mondada 2008, la transcription dans la perspective de la linguistique interactionnelle, in BILGER, Données orales les enjeux de la transcription, p81.
[94] Gadet 2008, "L'oreille et l'œil à l'écoute du social", in Bilger Données orales, les enjeux de la transcription, p44.
[95] Gadet 2008, "L'oreille et l'œil à l'écoute du social", in Bilger, Données orales, les enjeux de la transcription, p46

d'une transcription standard. Nous avons donc établi un niveau de transcription qui tienne compte de cette démarche réflexive et qui reste de la responsabilité du chercheur dans un cadre qu'il se doit d'expliciter.

Toutefois, afin de traiter un grand volume de données orales et de permettre la comparaison de segments, nous avons défini un autre niveau de transcription, appelé niveau zéro, dans le seul but de faciliter la navigation. En ce sens, la transcription est à considérer comme une simple annotation qui permet d'ajouter une information sur un segment.
Le niveau zéro (T0) s'appuie sur des conventions minimales selon les principes suivants :

- faciliter la navigation dans le signal (synchronisation pour réécoute),
- transcrire tous les mots, y compris les amorces de mots et les disfluences,
- définir un codage qui soit le plus explicite possible,
- déterminer un codage qui offre le minimum d'ambiguïté.

Ce niveau ne permet pas de produire des analyses. C'est un outil de préparation du corpus, la première des annotations.

Un grand corpus oral « disponible » : l'expérience du *Corpus d'Orléans* 1968-2012*. (Eshkol et al 2011,* 52 :26)

La transcription qui est le premier degré d'annotation de l'oral est une étape primordiale. C'est sur ce premier niveau que vont s'ajouter d'autres annotations. Les choix faits à ce stade influencent tout le traitement postérieur. La tâche a été d'autant plus difficile qu'il n'y a pas de conventions de transcription admises par la communauté scientifique.

Plusieurs contraintes ont influencé nos choix. On voulait en premier lieu mettre à disposition des chercheurs une grande quantité de données transcrites (700 heures d'enregistrement). Le processus de transcription devait donc être effectué rapidement mais avec une bonne efficacité. Il n'existe pas aujourd'hui d'outil de transcription automatique disponible, il s'agit donc de transcription manuelle. Ceux qui ont travaillé sur l'annotation manuelle savent que moins on annote d'informations, plus on gagne dans la quantité et la qualité car l'annotateur est moins dispersé et donc plus concentré sur sa tâche. Nous sommes allés dans la même direction et nous avons choisi l'annotation minimale. Il s'agit de la transcription orthographique qui conserve les spécificités de l'oral (amorces, disfluences, répétitions, etc.).

Du Français Fondamental aux ESLO (Abouda & Baude, 2008 :141)

Nous avons conçu cette première transcription à un degré le plus proche du zéro, en lui donnant uniquement le statut d'outil de navigation au sein du corpus sonore. L'outil sélectionné a été Transcriber pour sa simplicité d'utilisation, sa robustesse face à des fichiers longs, et sa sortie en un format de fichier XML qui nous a semblé être une garantie d'interopérabilité.

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

La T0 :

- La T0 est nécessairement imparfaite (la règle prévue dans ESLO est qu'un transcripteur ne s'arrête pas plus de 30 secondes sur un point problématique, il doit alors prendre une décision).
- la T0 implique nécessairement différentes versions modificatrices (V0.1 (V0.n)).
- la T0 doit permettre une lecture et une édition faciles (proche des usages de l'écrit).
- la T0 doit correspondre aux normes des outils du traitement de corpus écrits (impliquant le minimum de prétraitement du corpus cf. Valli & Veronis 1999).

A cette T0, correspond une T1 qui est le second niveau dédié à l'analyse :
Le niveau T1 correspond à une transcription utilisable pour une analyse linguistique ayant des objectifs précis :

- finesse de transcription (corrections, choix théoriques),
- ajouts de codages spécifiques (prosodie, multitranscription …),
- ajouts d'autres annotations sur différents niveaux empilables.

Il y a une interdépendance entre les différents niveaux :

- Possibilité d'ajouter du codage de la T0 vers la T1,
- Possibilité de soustraire du codage de la T1 à la T0.

Cette interdépendance ne peut se construire qu'à partir d'un travail de standardisation et d'explicitation du codage (comme le permet la TEI).

Cette solution, opportuniste dans le cadre d'une mise à disposition d'un grand corpus oral, supposait la disponibilité d'une technologie de synchronisation du son et du texte pour un corpus numérique et l'élaboration de conventions minimales.

3.5.3. Transcription synchronisée [retour]

[http://cocoon.huma-num.fr/exist/crdo/display/crdo-ESLO2_ENT_1001_C?plugin=html5]

[http://modyco.inist.fr/data/eslo/eslo2/ESLO2_ENT_1001/ESLO2_ENT_1001_audio.html]

Les recherches en linguistique ont toujours été fortement influencées par l'outillage technologique. Nous l'avons précisé, la possibilité d'enregistrer la parole a représenté une avancée dont les perspectives sont considérables. De même, la possibilité de synchroniser le son et le texte au sein des corpus numérique modifie la relation du chercheur à son objet.

C'est vrai en particulier pour la transcription qui peut accéder à une forme d'annotation fondée sur la relation entre l'information ajoutée et le segment isolé.

Depuis le début des années 2000, plusieurs logiciels d'aide à la transcription synchronisée ont été développés. En 2004, année de démarrage des transcriptions, nous nous sommes appuyés sur la réflexion en cours au sein d'une communauté naissante à laquelle participait l'équipe ESLO pour choisir un logiciel de transcription :

	PRAAT	ELAN	TRANSCRIBER
Site web	http://www.fon.hum.uva.nl/praat	http://www.latmpi.eu/tools/elan	http://sourceforge.net/projects/trans/files/TRANSCRIBER/
Auteurs	Paul Boersma David Weenink	Birgit Hellwig (auteur original) Dieter Van Uytvanck Micha Hulsbosch (auteurs des mises à jour)	Karim Boudahmane Mathieu Manta Fabien Antoine Sylvain Galliano Claude Barras
Disponibilité	Accès libre	Accès libre	Accès libre
Plates formes	Windows, Unix, Macintosh	Windows, Unix, Macintosh	Windows, Unix, Macintosh
Format d'entrée	wav, mp3	Wav	.aif, .aiff, .au, .mp3, .ogg, .sd, .smp, .snd, .sph, .wav
Format de sortie	Textgrid	Shoebox > elan, chat > elan, Transcriber > elan	.html, .lbl, .stm, .txt, .typ

Tableau extrait de la thèse en cours de L. Hriba (inédit)

Après différents tests, y compris des essais de transcription automatique dans le cadre d'une collaboration avec le LIMSI, le choix s'est porté sur le logiciel Transcriber.

Avantages :

- ✓ synchronisation enregistrement-transcription,
- ✓ fichier à structure XML,
- ✓ marquage des tours de parole, de la segmentation thématique des conditions acoustiques,

- ✓ balises des évènements : bruits, lexique, prononciation, langue…
- ✓ interface simple avec une prise en main rapide,
- ✓ robustesse.

Inconvénients :

- ✓ difficulté de gestion des chevauchements,
- ✓ choix réduits par la DTD.

3.5.4 Conventions et process [retour]

Le choix des conventions de transcription a donné lieu à un travail du même ordre. En effet, de manière surprenante, cette tâche est moins simple qu'il n'y parait puisque même si les travaux de Claire Blanche-Benveniste sur la transcription du français parlé ont permis une avancée considérable (Blanche-Benveniste & Jeanjean 1987)[96]. Il existe peu d'informations précises sur les conventions de transcription utilisées par les linguistes dans le cadre de leurs projets de recherche. La situation évolue mais elle était particulièrement visible en 2004.

Fort du constat qu'il n'existait pas un standard, nous avons réalisé une étude à partir des conventions disponibles à partir de cinq projets :

- VALIBEL : VAriétés LInguistiques du français en BELgique
- TRANSCRIBER (GGA-LIMSI) : Guide d'annotation
- CLAPI (ICAR) : Corpus de Langue Parlée en Interaction
- DELIC : Description Linguistique Informatisée sur Corpus
- PFC : Phonologie du Français Contemporain

Cette analyse a donné lieu à un tableau détaillé dont nous pouvons ici fournir les éléments principaux.

Il existe tout d'abord des principes partagés :

- Transcription orthographique.
- Absence de ponctuation (excepté le point d'interrogation pour certains).
- Segmentation en tour de parole (mais différence de codage).
- Codage de certains phénomènes spécifiques de l'oral (disfluences, etc.).
- Codage des locuteurs.

Ces principes correspondent à des convergences théoriques :

- Lisibilité (orthographe standard non aménagée).
- Spécificité de l'objet "forme orale du français" :
  - pas de ponctuation,
  - marquage des disfluences,
  - segmentation en tour de parole,
  - codage du locuteur (anonymisation).

[96] Blanche-Benveniste, C., & Jeanjean, C. (1987). *Le français parlé: transcription et édition*.

- Volonté d'un minimum d'interopérabilité par un codage explicite et structuré

Il existe aussi des différences dans la sélection des éléments codés :

	TRANSCRIBER	DELIC	VALIBEL	CLAPI
Allongement	X	elle est jolie:	X	elle est jolie:
Intonation	X	X	X	COR : tu y vas/ toi\ COR : tu y vas//
Liaisons	Liaisons erronées : vingt+[pron=vingt-z] animaux	Absence de liaison remarquable : Ex : plusieurs # éléments Présence de liaison Ex : qui=z=ont	SAMPA : liaison remarquable	X
Elisions	X	X	X	Non standard : COR : il nous faut d`jà ça

Et dans les choix de codage :

	TRANSCRIBER	DELIC	VALIBEL	CLAPI
Pauses		Pause brève : + Pause longue : ++	Pause brève : / Pause longue : //. Pause très longue : (silence)	chronométré à l'aide d'un logiciel au 10ème de seconde. (2.2)
Passages inaudibles	Balise prononciation : inintelligible ou inintelligible faible	syllabe inaudible * suite de syllabes inaudibles : **	Parenthèses (x) = une syllabe et (xx) = un groupe de syllabes et (xxx)= un passage plus long	segment inaudible (xxx) et (inaud.) lorsque complètement inaudible
Multi-transcriptions		/des , deux/	{des , deux }	(des , deux)
Amorces	après-de(main) ou il faut les rem() les remplacer	il faut les rem- les remplacer	il faut les rem/ les remplacer	il faut les rem- les remplacer

Les différences de codage peuvent relever d'objectifs différents mais explicites : analyse de la conversation, phonologie, français parlé etc.

Voici un exemple de codage des chevauchements dans le projet CLAPI :

Insertion de crochets "[" et "]", encadrant le chevauchement dans chaque tour. Les crochets ouvrant "[" (début du chevauchement) sont obligatoires (sans espace après) ; les crochets fermants "]" (fin du chevauchement) par contre sont facultatifs. Les crochets sont alignés verticalement au moyen d'espaces (attention : ne pas utiliser tabulation).	SAR A:::ïe COR c'est pas utile [ça] SAR [ça] mar[che] COR [attr]ape le SAR c'est bon/ SAR salut corinne COR sa[lut] SAR [tu] sais ce [qu'ils ont dit/] MME [bonjour sarah] SAR bonjour madame

Mais ces différences peuvent aussi relever de choix implicites.

Un exemple : le codage des pauses

DELIC : distinction pause brève, pause longue sans indication de ce qui les différencie.

CLAPI : le terme de pause est réservé à des "pauses intra-tour", le terme de "silence" est réservé au tour de parole sans production (ils peuvent être non attribués à un locuteur).

Silence à valeur de tour	Les silences sont chronométrés à l'aide d'un logiciel au 10ième de seconde près, sauf pour les silences d'une durée inférieure à 0,2 secondes qui sont notés par (.)	(0.7) (2.2) (.)
	Note : Dans le cas où le silence est attribuable à un participant, il est noté dans un paragraphe portant l'identifiant du participant concerné.	SAR c'est toi qui m'as piqué ma gomme/ COR (1.2) SAR mais ça va PAS ça

C'est à partir de ce travail que nous avons établi les conventions ESLO qui ont été rédigées afin de servir de support méthodologique de référence. C'est un document à part entière du corpus.

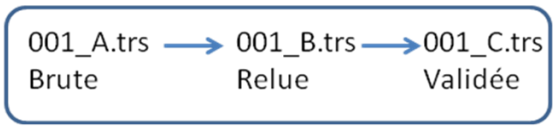
Les grands principes sont les suivants :

- Transcription synchronisée / XML.
- Choix de l'orthographe standard non aménagée.
- Codage minimal.
- Pas de codage ambigu.
- Tout élément sonore est transcrit (sous sa forme orthographique).
- Aucun élément n'est ajouté.
- Pas de ponctuation (exception ?).

- Segmentation intuitive (segments courts : groupe de souffle et/ou unité syntaxique).
- Pause = segment vide.
- Majuscule.
- Dictionnaire de référence + lexique.
- Usage réduit des balises.

Réalisation d'une triple transcription donnant lieu à 3 versions :

* Version A : transcription brute réalisée le plus rapidement possible,
* Version B : relecture par un autre transcripteur,
* Version C : correction par un relecteur confirmé.



Temps calculé:

- Version A: 10 fois le temps
- Version B: 5 fois le temps
- Version C: 5 fois le temps

Une heure d'enregistrement ESLO correspond donc à 20h de travail pour obtenir une version finalisée.

Les conventions ont été intégrées à un guide du transcripteur car l'originalité du projet de transcription des ESLOs réside dans l'élaboration et la maintenance d'une procédure permettant une transcription par de nombreux contributeurs[97] en flux tendu.

Phase de transcription des ESLO

Version 1, mai 2008 http://www.nakala.fr/data/11280/caea0537

Une première version de vingt-cinq pages a été rédigée dans le cadre des travaux de la thèse de Linda Hriba. On en recopie le plan:

[97] Précisons que le corpus ESLO a été transcrit par des étudiants vacataires. Il s'agit en fait exclusivement d'étudiantes recrutées parmi les promotions de licence et de master en sciences de langage de l'université d'Orléans. Dans la suite du document, l'usage du féminin sera donc systématique pour la présentation de ces étudiantes et de leur travail.

Version 2, février 2010 : http://www.nakala.fr/data/11280/7445014a

Intitulé « guide du transcripteur et du relecteur », cette deuxième version, réalisée à partir de la V1, a été rédigée principalement par Céline Dugua qui venait de rejoindre l'équipe et qui a pris en charge ces tâches dans le cadre d'une coresponsabilité de gestion du projet ESLO.

Quatre grands types de modification

- 1) Affinement de **la méthode à appliquer pour la transcription** (à l'usage des transcriptrices)
    - Rajout d'une partie « outils d'aide à la transcription et à l'organisation ». Objectif : proposer des outils aux transcriptrices pour améliorer la communication interne avec le souci de conserver les informations/les décisions prises.
        - **Création d'une base de données** : « *Une base de données permet de réunir l'ensemble des informations sur les enregistrements, les*

locuteurs et les transcriptions ESLO2. Vous pourrez vous y reporter pour obtenir notamment les codes des locuteurs. »

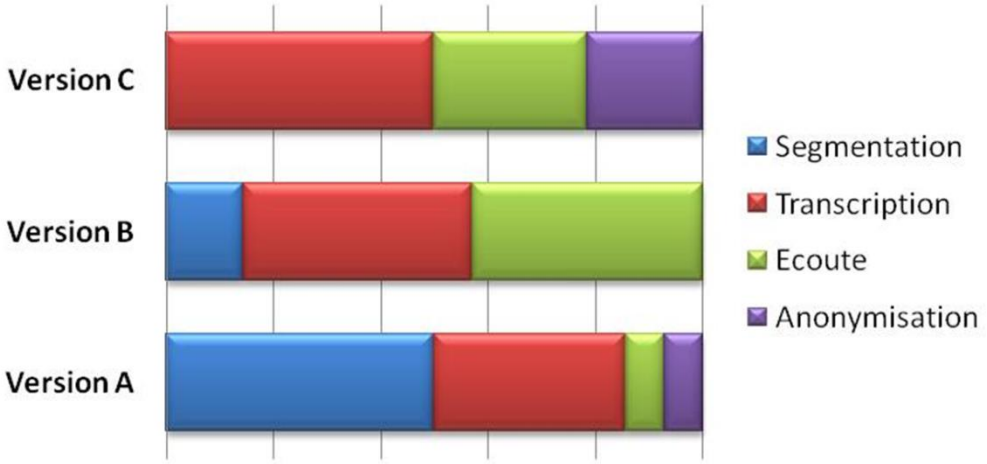
- **Mise en place d'un Google Groupe** : Un « Google Groupe Transcription ESLO » a été créé afin de faciliter la communication entre les transcriptrices et les chercheurs, et entre les transcriptrices entre elles. » Ce forum permettait de poser des questions et de mutualiser les réponses en y intégrant le dépôt des documents pertinents pour la tâche entreprise.

- Des précisions sur le logiciel ont été apportées concernant par exemple les paramétrages à faire au début, l'installation d'un dictionnaire sur le poste de travail, etc.
- Rajout d'une partie « Etapes à suivre pour faire une transcription », en 6 étapes (p.11 du guide V2). Objectif : cadrer, homogénéiser les pratiques des transcriptrices

- 2) Ajout d'une partie sur la **relecture** (versions B et C) où sont précisées les 3 étapes concernant les tâches à accomplir.
- 3) Suite à un travail collectif des membres de l'équipe, **un ajustement des conventions.**
- 4) Un travail sur la **mise en page du document**

Version 3, juin 2011 **http://www.nakala.fr/data/11280/96887cc0**

La principale modification concerne l'ajout d'une partie « Procédures de transcription, relecture, validation » afin de recenser les tâches inhérentes à la transcription et à la relecture pour en répartir l'application entre les 3 versions. Il s'agit de limiter le nombre de tâches dévolues à chaque version en garantissant l'exécution des procédures de vérification. Voici un schéma récapitulatif de l'ensemble des tâches pour chaque version.



Graphique des tâches de transcription[98]

Binômes transcriptrice/relectrice : A partir de la V3, les binômes mis en place sont dissociés pour deux raisons :

- Organisation : on a moins de couples A/B stables, soit que des versions B ne soient pas effectuées en continuité, soit que les transcriptrices aient des contrats différents en nombre d'heures.
- Par rapport au protocole de transcription/relecture, la stabilisation de binômes induit des pratiques de routinisation et certaines « erreurs » cumulées, que commettent également A et B, ne sont pas relevées.

Il existe d'autres modifications dictées par les réquisits de la base de données. Ainsi en cas d'hésitation d'attribution entre deux locuteurs, jusqu'à la V2 du guide, il fallait porter un point d'interrogation ce qui, dans la suite du process de gestion de la BDD, revenait à créer un nouveau locuteur. A partir de la V3, la consigne a été d'opérer systématiquement un choix et un plan de nommage des locuteurs a été élaboré et rajouté dans le guide.

Nous avons modifié le guide en tenant compte des remarques et des questions des transcriptrices et de nouvelles formes présentes dans ESLO2 ont nécessité la mise en place d'un « Lexique ESLO ».

> Guide V3, p. 5 : « *Par ailleurs, ce guide est accompagné d'un document* Lexique-Eslo *qui recense un certain nombre de graphies particulières, de difficultés orthographiques, etc. Ce* Lexique*, fourni en version électronique, est mis à jour*

[98] Tableau et analyse réalisés par Céline Dugua.

régulièrement. Dès qu'une modification est apportée, nous vous envoyons la dernière version par mail. »

Version 4, mai 2013 http://www.nakala.fr/data/11280/2cf7a33a
Le dictionnaire de référence jusqu'à la V3 était *Le Robert*. A partir de la V4, la BU d'Orléans n'étant plus abonnée au *Robert en ligne*, le changement s'est fait au profit du *TLFi*.

Cette version a été de surcroît complétée par une « fiche remarque » confiée aux transcriptrices afin qu'elles notent toutes leurs observations concernant leur travail de transcription.

La phase de transcription est l'une des plus fastidieuses qui soit. Elle a nécessité l'engagement d'une grande partie de l'équipe ESLO (principalement Olivier Baude, Céline Dugua, Layal Kanaan-Caillol, Linda Hriba, Gabriel Bergounioux) et la contractualisation de 45 étudiantes-vacataires. Elle a entrainé différentes collaborations :

- Tests de transcription automatique avec le LIMSI.

- Réalisation d'un transcodage en TEI (TEIML) dans le cadre d'Ortolang.

- Collaboration à un nouvel outil (Transcriber-JS) réalisé par Christophe Parisse.

- Process d'anonymisation de la transcription.

Ces projets sont détaillés *infra*.

L'apport le plus innovant résulte de la démarche réflexive adoptée et de l'interdépendance d'une pratique du corpus et de la conduite d'analyses. On en lira la synthèse dans les deux chapitres suivants.

TEI et TEIML
http://www.tei-c.org/index.xml

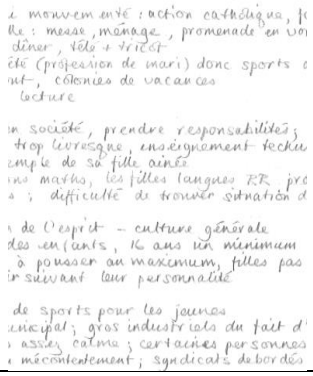
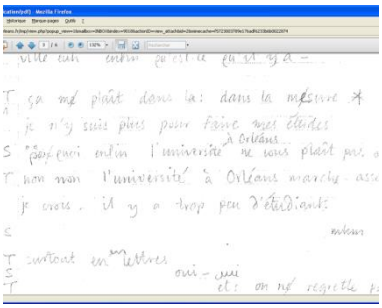
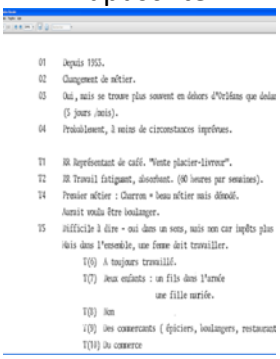
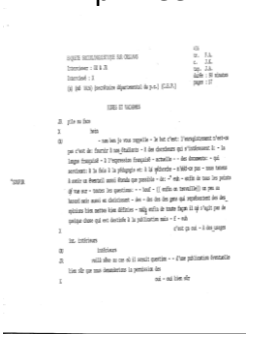
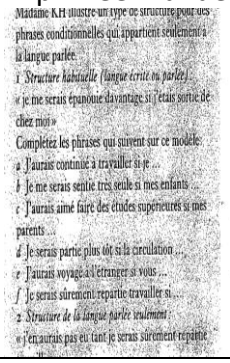
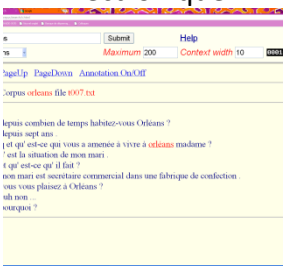
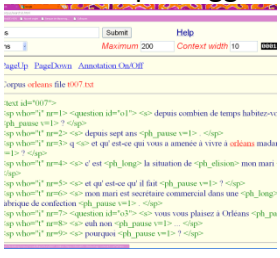
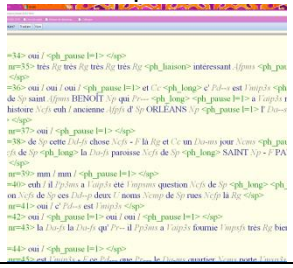
http://modyco.inist.fr/sources/Ortolang/DescriptionFormatTEI.pdf

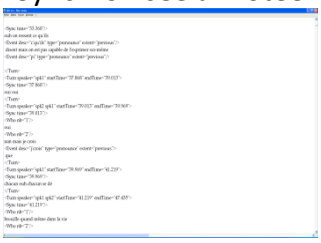
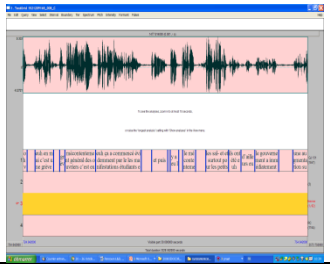
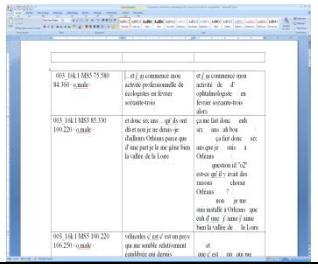
Il faut noter que l'objectif d'interopérabilité atteint a permis la conversion aisée des transcriptions ESLO au format TEIML (TEI adaptée à l'oral) dans le cadre des traavux de l'EQUIPEX ORTOLANG. Le format teiml est un format XML qui se base sur la norme TEI pour représenter la transcription du langage oral. Ce format a été développé à partir de la conversion de transcriptions aux formats Chat et Transcriber. Un document TEIML est

composé de deux éléments principaux : • teiHeader : contient les informations relatives à l'enregistrement • text : contient la transcription de l'enregistrement.

3.6.5 La transcription : quel objet ? [retour]

Le recul de quarante ans d'exploitation du corpus ESLO permet de suivre l'évolution d'un objet aussi particulier que l'est la transcription. Entre 1968 et 2015 la forme écrite d'un enregistrement ESLO a pris douze formes différentes :

Manuscrite	Manuscrite avec codage	Tapuscrite
		
Imprimée	Imprimée Manuel	Electronique
		
Electronique taggée	Electronique annotée	Trancriber
		

		
Synchronisée annotée 	Praat 	Doc 

Encore ne s'agit-il pas d'un inventaire exhaustif.

Comment décrire un objet aussi polymorphe ? L'expérience d'ESLO permet de proposer quelques pistes fondées sur le travail empirique :

- Par les métadonnées
  Celles-ci permettent de déterminer un titre, une date, un auteur et livrent différentes informations. La question est celle de la définition d'un format des métadonnées dont l'objectif peut différer selon qu'on se réfère à des opérations de conservation ou d'exploitation scientifique, en référence notamment à un cadre théorique donné. Ainsi, selon les domaines, le locuteur peut être dénommé « témoin », « participant », « contributeur »… La typologie des données et les différentes catégories mobilisées varie selon les projets.

- Par les cadres théoriques du corpus
  Ceux-ci sont rarement explicites dès l'origine du projet et, lorsqu'ils existent, il arrive fréquemment qu'ils n'aient pas été conservés. Ils sont donc peu accessible pour des études diachroniques.

- Par la définition des éléments constitutifs du corpus, leur structure et la granularité choisie :
  - Structure interactionnelle
  - Contenu général
  - Sens de l'énoncé (*bona fide*)
  - Gestes / regards / proxémique
  - Structure discursive

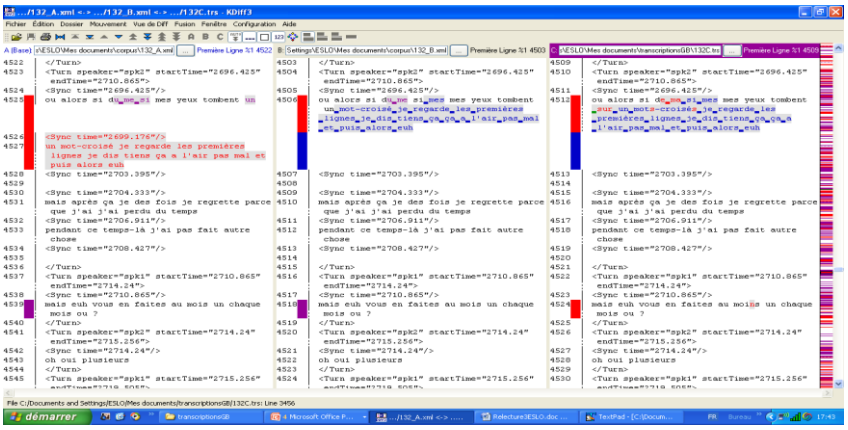
  - Structure syntaxique
  - Annotation (tags morphosyntaxiques)
  - Annotations (diverses)
  - Verbal
  - Verbal extensif (disfluences)
  - Phonétique
  - Éléments suprasegmentaux
  - Codage phonologique
  - …
- Par les outils, instruments et conventions utilisés

Tout ceci plaide en faveur d'une science du corpus qui se donne les moyens de définir son objet, ce qui implique de concevoir à la fois des projets et des données « situées ». Dans l'état actuel, force est de constater que les conditions de définition de l'objet constituent intrinsèquement un des facteurs de « variations » induites ou non contrôlées mais dont les effets sont bien visibles dans les analyses subséquentes.

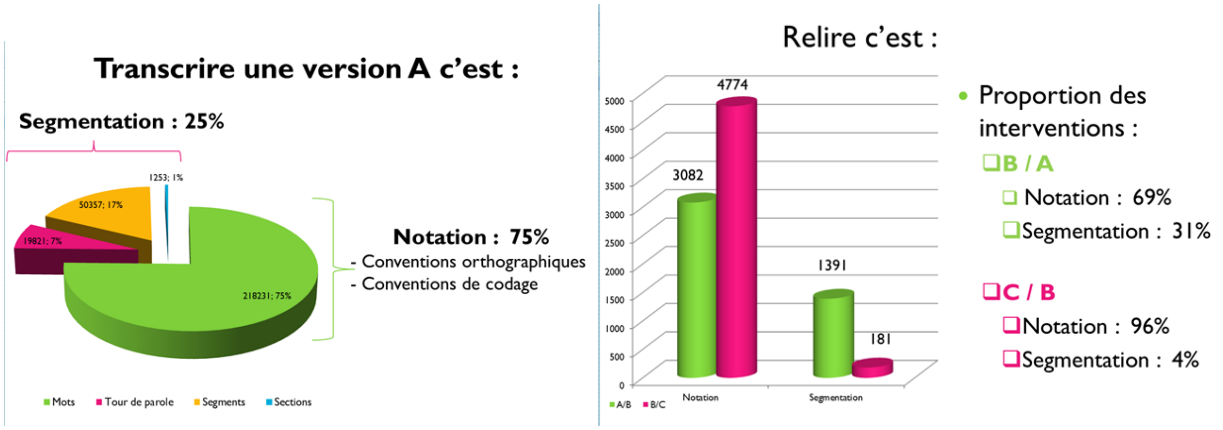
3.5.6. Transcription et variation [retour]

Ce chapitre présente certains éléments de la thèse (inédite) de Linda Hriba. La phase de transcription d'ESLO a produit un matériau d'un nouveau type qui n'a pas d'équivalent en français : les trois versions de transcriptions d'un grand corpus oral de français parlé. L'étude des conditions de production de ces transcriptions permet une approche inattendue de la variation linguistique et ouvre des perspectives pour des analyses en sociolinguistique et psycholinguistique.

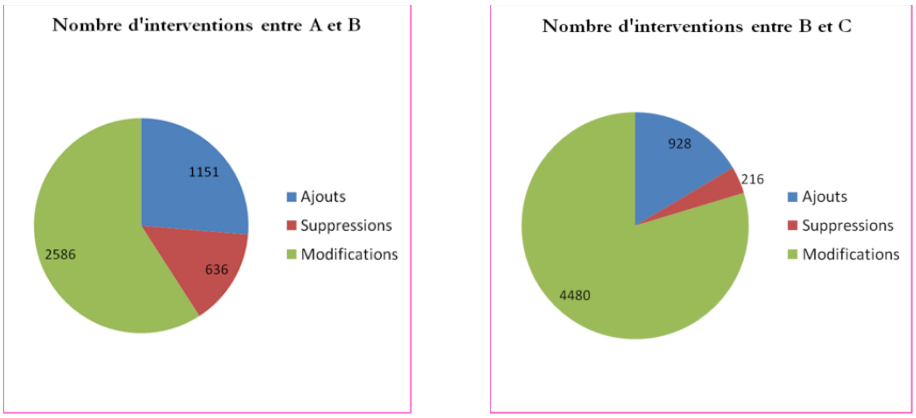
L. Hriba a réalisé un sous-corpus d'études afin de comparer les trois versions de transcription. Dans un premier temps, cette comparaison a été réalisée à l'aide d'un logiciel de comparaison de fichiers XML (KDIFF3).

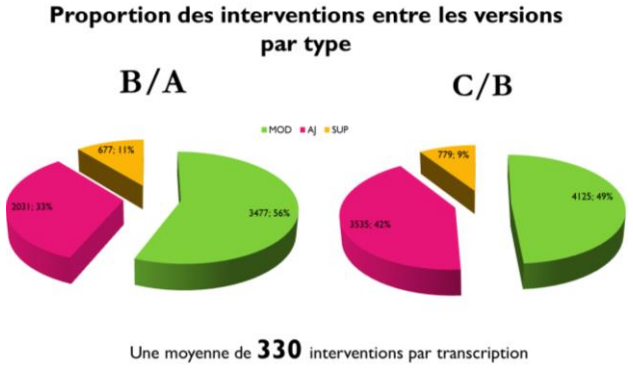


Ainsi est-il possible de décrire plus finement ce que représente la tâche de transcription au sein du process :



Une première analyse permet de repérer trois types principaux de modifications :





Nous pouvons dans un premier temps relever l'ampleur des interventions : plus de 300 pour une heure de parole.

Nous constatons qu'il y a plus d'éléments « oubliés » qu' « ajoutés » dans la version A. (26% d'ajouts entre A et B pour 16% de suppressions – 15% d'ajouts entre A et B pour 4% de suppressions). Les éléments oubliés concernent majoritairement :

- Les répétitions.
- Les onomatopées et interjections .
- D'autres éléments (séquences verbales, déterminants, prépositions…).

Les interventions concernent principalement des modifications (56% entre A et B et 49% entre B et C) qui se répartissent en trois catégories :

- La segmentation : les tours de parole, les pauses, les chevauchements.
- Les rectifications d'orthographe ou de transcription.
- « Les différences d'écoute » : élucidation de passages incompris, inversions, distorsions.

Cette dernière catégorie est la plus intéressante. Ces différences de perception peuvent intervenir dans un contexte de mauvaise qualité du signal, due à la qualité acoustique de l'enregistrement ou à la présence de bruits parasites. Mais il y a aussi des différences de perception alors que la qualité du signal est satisfaisante. L'étude permet de repérer des variations dues aux contextes et aux savoirs mobilisés par la transcriptrice. Elle analyse et interprète afin de reconstruire un énoncé au sens « plausible » selon le contexte ou une « norme syntaxique » (où jouent un rôle les préjugés sur l'oral). Elle anticipe ce qui va être dit (ce qui pourrait être dit) à partir de la connaissance qu'elle a du locuteur et de « routines linguistiques» relevant de son propre habitus linguistique :

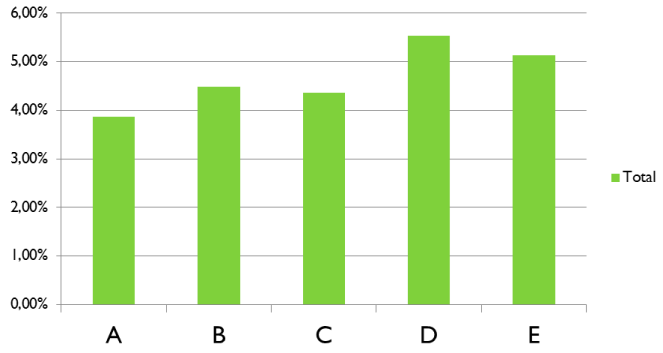
Exemples :

062_B	062_C
en **tapis noir**	en **tapinois**

109_B	109_C
y a personne d'autre qui **aboie** aux chapitres je pense	y a personne d'autre qui **a voix** au chapitre je pense

107_B	107_C
on restera	**nous** resterons

La possibilité de croiser ces informations avec la catégorisation sociale des locuteurs établie par Alix Mullineaux montre que ces « mauvaises » modifications sont significativement plus fréquentes quand le transcripteur écoute un locuteur d'une catégorie sociale D/E par rapport à un locuteur d'une catégorie ABC :



On peut ainsi concevoir la transcription comme une formidable porte d'entrée sur la langue dans la conception variationniste de celle-ci. En effet ce qu'entend (et ce que n'entend pas), ce qu'écrit un transcripteur est le produit de conditions sociales et d'ajustement de cette production à un marché linguistique. Nous suivrons Encrevé dans l'importance qu'il accorde à la relation auditeur/locuteur comme lieu de la compétence qu'intériorise et extériorise le sujet :

> *« Car c'est dans le jeu des grammaticalités au sein même de la langue que doit se saisir le rapport auditeur-locuteur et la forme des compétences qu'intériorise et extériorise le sujet. Ainsi peut-elle s'interroger sur l'intériorisation de la structure sociale inhérente à l'intériorisation de la structure sociale inhérente à l'intériorisation du savoir linguistique et qui se traduit par la différence entre audition et locution : ce qui est entendu et n'est pas (re)produit. Quelles sont les propriétés sociales par rapport à celles du sujet en question ? La langue du locuteur c'est celle des groupes des égaux, sa langue d'auditeur inclut celles des autres (supérieurs et /ou inférieurs) ; les variations régulières représentent les effets des grammaires des « autres », qu'elles soient stylistiques (produites par les modifications du contexte immédiat) ou inhérentes (dans un contexte inchangé), traces*

> *qui intériorisent à leur manière le statut sociale relatif des grammaticalités ».* (Encrevé 1977 :11[99])

L'étude des variations dans les transcriptions du corpus ESLO nous permet donc de ne pas réduire ce matériel à celui issu d'une phase de préparation préalable à l'analyse, la transcription comme « donnée du linguiste » étant aussi le lieu de variations qui révèlent la langue et sa « nature sociale ».

Ainsi, la phase de transcription, si elle permet d'appréhender la variation, est à son tour source de variations :

- Les variations issues de la place du programme au sein d'un champ scientifique. Ce point est repérable dans la présentation des conventions de transcriptions qui, par-delà la recherche d'un formalisme, sont partie prise et partie prenante d'une méthodologie au sein d'un champ traversé par des luttes et des enjeux symboliques. Ces enjeux peuvent concerner une position théorique mais aussi la place d'un laboratoire, notamment lors de la mise en compétition lors d'appels d'offres et dans la recherche de financements.

- Les variations dues aux contraintes technologiques.

- Les variations du transcripteur comme agent d'un fait social.

[99] ENCREVE, P. (1977). « Présentation : Linguistique et socio-linguistique ».

3.6 Gestion, conservation, accès et diffusion du corpus

Articles et livre :

- 2006, *Corpus oraux, Guide des bonnes pratiques*, https://halshs.archives-ouvertes.fr/halshs-00355472
- 2008, Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO, https://halshs.archives-ouvertes.fr/halshs-01162506
- 2009, Un grand corpus oral « disponible » : le corpus d'Orléans, 1968-2012, https://halshs.archives-ouvertes.fr/halshs-01163053
- 2014, « Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO », https://halshs.archives-ouvertes.fr/halshs-01165957
- 2015, « Les ESLOs, du portrait sonore au paysage digital », https://halshs.archives-ouvertes.fr/halshs-01165907

Communications orales :

- 2006, Interoperability of audio corpora : the case of the french corpora, https://halshs.archives-ouvertes.fr/halshs-01162927
- 2006, Constitution et exploitation d'un grand corpus de "données situées" Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), https://halshs.archives-ouvertes.fr/halshs-01165954
- 2008, Un grand corpus de référence du français parlé : état des lieux et perspectives, https://halshs.archives-ouvertes.fr/halshs-01165952, https://halshs.archives-ouvertes.fr/halshs-01165954
- 2008, Les enquêtes sociolinguistiques à Orléans, Base et corpus, https://halshs.archives-ouvertes.fr/halshs-01165996
- 2014, Le corpus des ESLO à l'ère des Digital Humanities, https://halshs.archives-ouvertes.fr/halshs-01165909,
- 2015, Archivage de corpus oraux : Etat des lieux à partir de l'exemple du corpus ESLO, https://halshs.archives-ouvertes.fr/halshs-01165908
- 2015, Mutualiser et diffuser des corpus : vraiment ? Pourquoi ? Comment ?, https://halshs.archives-ouvertes.fr/halshs-01165906

Documents et sitographie :

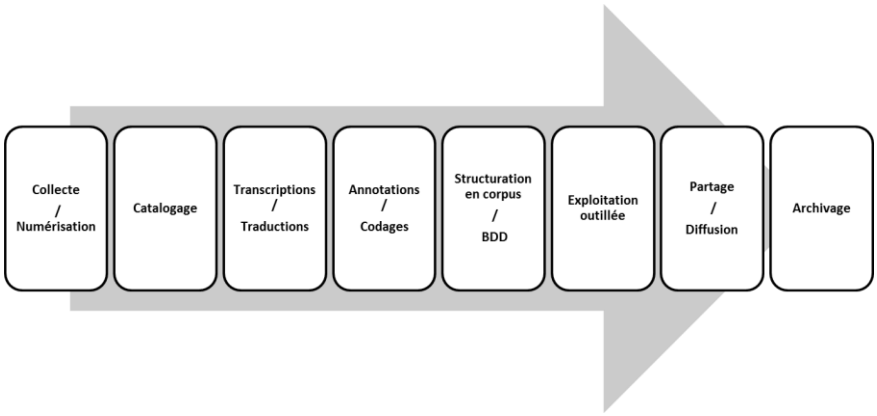
- Site ESLO : http://eslo.huma-num.fr/
- Cocoon : http://cocoon.huma-num.fr/exist/crdo/collection_eslo.htm
- Ortolang : https://www.ortolang.fr/#/market/corpora/c2f06cb1-6458-4699-81e2-7f46a94a3e4f
- BnF : http://archivesetmanuscrits.bnf.fr/ead.html?id=FRBNFEAD000095934
- Guide d'anonymisation : https://www.nakala.fr/nakala/data/11280/9ce4049c

La forme relativement stable que connait le corpus ESLOs actuellement provient sans nul doute des opérations de conservation, d'archivage et de diffusion de celui-ci. Même s'il faut différencier ESLO1 qui est un corpus clos, c'est-à-dire relativement figé, et ESLO2 qui est un corpus ouvert, tous deux sont sensibles aux effets d'une exploitation outillée. Là encore, l'évolution du traitement des ESLOs a été accompagnée par les actions nationales et européennes en termes d'infrastructures de recherche, de conservation et de mutualisation des données de la recherche.

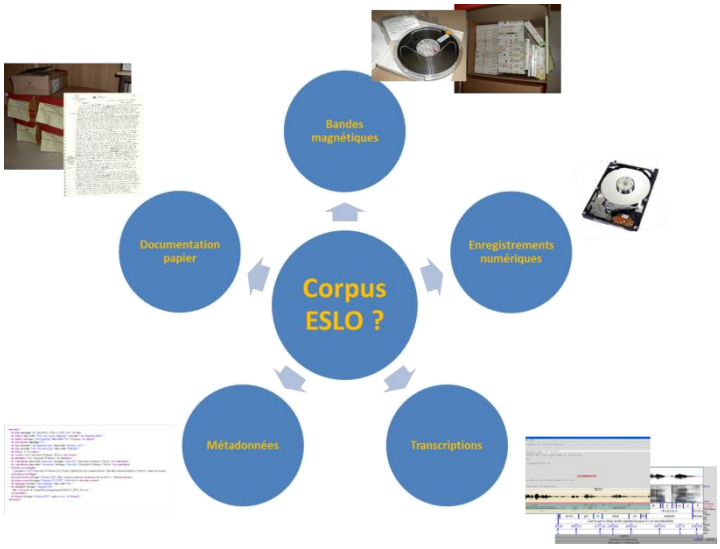
Les corpus des ESLOs ont donné lieu à un process de gestion des données afin de permettre leur conservation, leur diffusion et leur exploitation. Ce process, les choix et les outils développés l'ont été de manière ad hoc car au démarrage du projet aucun de ceux-ci n'étaient disponibles. Il s'agit donc, à travers cette expérience, de revenir sur une micro histoire du tournant numérique des corpus oraux en linguistique.

3.6.1 Process général [corpus de la parole] [retour]

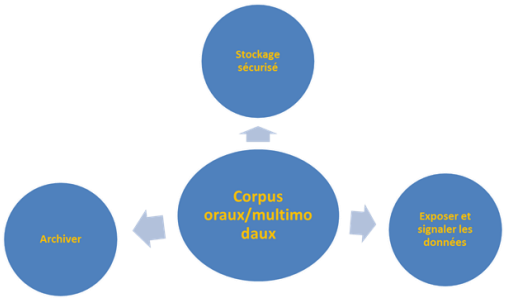
Le process général peut être schématisé de la sorte :



Dans le cas du corpus ESLO, celui-ci est composé de bandes magnétiques (ESLO1), d'enregistrements numériques (ESLO2 et ESLO1 numérisé), de transcriptions, de métadonnées et de documentation papier.



C'est l'ensemble de ces objets dont il va falloir assurer la description, la conservation et l'accès afin de répondre au triple objectif de toute donnée de la recherche dans le cadre des humanités numériques : assurer le stockage, exposer et signaler les données et permettre l'archivage.

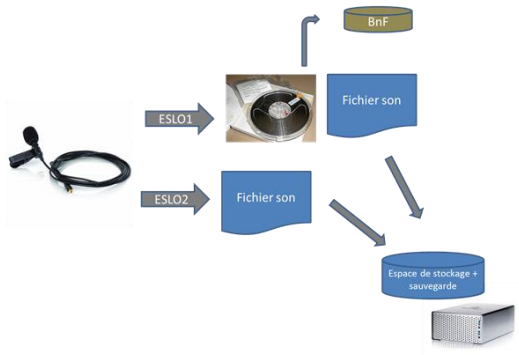


En réponse à ce triple objectif, nous avons organisé un process à partir du cycle de gestion d'un document ESLO.

3.7.2 De la collecte au site « ESLO » [retour]

Collecte des documents sonores

La base du corpus ESLO, ce sont les documents sonores. Ainsi, le cycle de vie d'un document du corpus ESLO commence par la phase de collecte, soit par la numérisation, soit par captation numérique.



Description et catalogage [corpus de la parole]

Les enregistrements sont ensuite transcrits. Il faut alors stocker ces données, en assurer la conservation, les décrire et y donner accès. Ces différentes opérations se font à l'aide de l'application-WEB ESLO développée par la société GFI. Cette application et les données ESLO sont hébergées sur la grille Huma-Num (CC In2p3) qui en assure le stockage sécurisé.

Cette application concentre donc un module de gestion des documents, un module d'archivage et un site de diffusion.

Ces trois modules sont interdépendants. La gestion des documents sous la forme de formulaires afin d'alimenter une BDD reprend les éléments utiles à l'archivage (notamment la description) et ce sont également ceux-ci qui seront utilisés par les outils de requêtes liés à la diffusion.

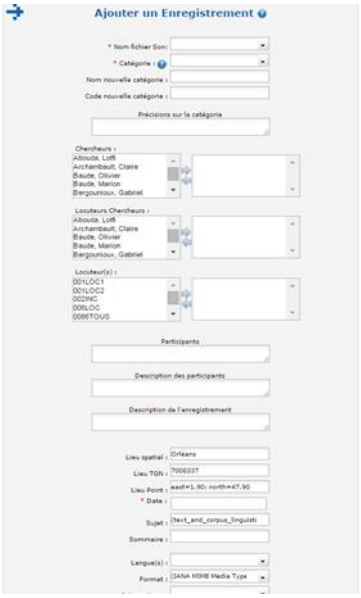
Le document sonore et sa transcription deviennent les deux éléments nécessitant une description. Nous distinguons donc les métadonnées décrivant les enregistrements et les transcriptions et les métadonnées enrichissant la situation de production linguistique et notamment le profil sociologique du locuteur.

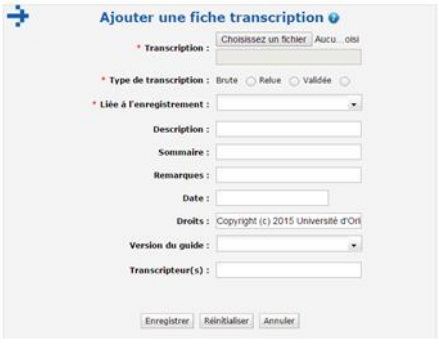
Chaque enregistrement (fichier sonore) et sa/ses transcription(s) (fichier de Transcriber) sont stockés dans une base de données où ils sont liés avec d'autres informations relatives au locuteur principal de chaque entretien : date et lieu de naissance, sexe, profession et appartenance sociale.

Ces informations sont renseignées au sein du module de gestion de l'application. Il s'agit d'un espace de contribution accessible pour les membres de l'équipe. Il permet de décrire les enregistrements et les transcriptions et de les déposer sur les serveurs de la grille Huma-Num à l'aide d'un protocole SFTP.

Espace de contribution de l'application WEB

La description se fait à l'aide de différents formulaires :



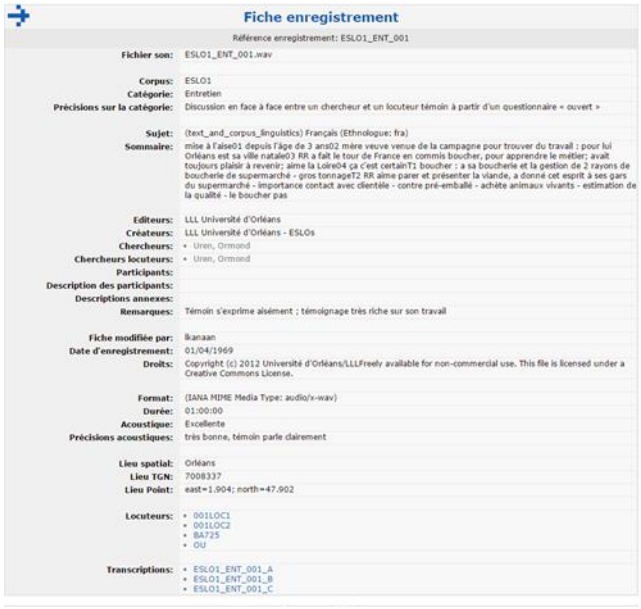


Les métadonnées sont celles requises par les opérations d'archivage plus celles nécessaires à l'exploitation scientifique par l'équipe du projet. Les métadonnées requises par la procédure d'archivage de l'entrepôt Cocoon (quinze étiquettes extraites du standard DUBLIN-CORE OLAC) sont les suivantes (DUBLIN CORE OLAC) :

Titre de l'enregistrement ESLO - bande 006
Description Extrait du corpus d'Orléans,
Lieu d'enregistrement France, Orléans: salle de séjour chez témoin
Lieu d'enregistrement 7008337
Lieu d'enregistrement east=1.90; north=47.90
Date de l'enregistrement 1969-04-03
Fichier de l'enregistrement
Type d'enregistrement primary_text
Type d'enregistrement interactive_discourse
Contributeur [depositor] Université d'Orléans/CORAL
Contributeur [interviewer] Université d'Orléans/CORAL
Contributeur [researcher] liste de noms
Contributeur [transcriber] M'Charek, Linda
Contributeur [transcriber] Petit, Mélanie
Contributeur [speaker] JK
Contributeur [speaker] FD 237
Contributeur [speaker] femme
Maison d'édition Université d'Orléans/CORAL
Droits d'auteurs Copyright (c) 1993 Université d'Orléans/CORAL
dc:creator Blanc, Michel , dc:creator Biggs, Patricia, dc:creator Ross, John, dc:creator Kay, Jack dc:creator Dalwood, Mary
dc:source Bande 006
dcterms:accessRights Freely available for non-commercial use
dcterms:license http://creativecommons.org/licenses/by-nc-nd/2.5/
dcterms:isRequiredBy oai:crdo.vjf.cnrs.fr:crdo-FRA_ESLO_omelette006
dcterms:extent PT1M55S

Ceci donne deux accès principaux aux métadonnées :

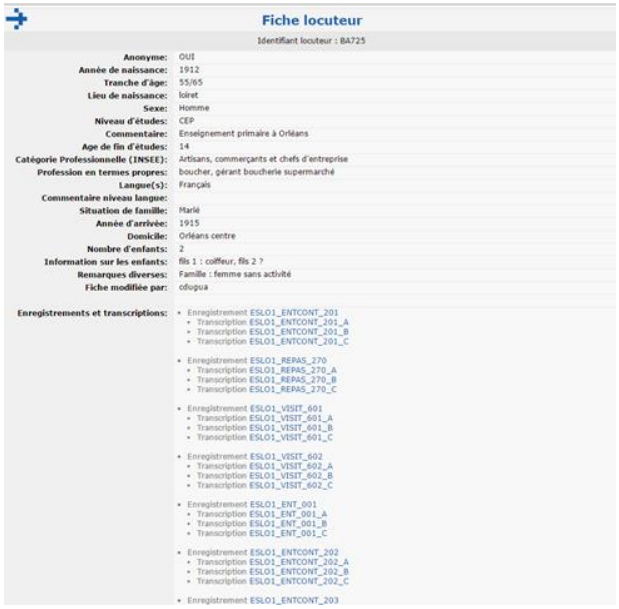
- soit par le site ESLO avec les choix de l'équipe pour une exploitation scientifique (notamment en intégrant les informations sur le locuteur comme la catégorisation en CSP et selon l'échelle AM :



Fiche enregistrement

Référence enregistrement: ESLO1_ENT_001

Fichier son:	ESLO1_ENT_001.wav
Corpus:	ESLO1
Catégorie:	Entretien
Précisions sur la catégorie:	Discussion en face à face entre un chercheur et un locuteur témoin à partir d'un questionnaire « ouvert »
Sujet:	(text_and_corpus_linguistics) Français (Ethnologue: fra)
Sommaire:	mise à l'aise01 depuis l'âge de 3 ans02 mère veuve venue de la campagne pour trouver du travail : pour lui Orléans est sa ville natale03 RR a fait le tour de France en commis boucher, pour apprendre le métier; avait toujours plaisir à revenir; aime la Loire04 ça c'est certainT1 boucher : a sa boucherie et la gestion de 2 ravons de boucherie de supermarché - gros tonnageT2 RR aime parer et présenter la viande, a donné cet esprit à ses gars du supermarché - importance contact avec clientèle - contre pré-emballé - achète animaux vivants - estimation de la qualité - le boucher pas
Editeurs:	LLL Université d'Orléans
Créateurs:	LLL Université d'Orléans - ESLOs
Chercheurs:	• Uren, Ormond
Chercheurs locuteurs:	• Uren, Ormond
Participants:	
Description des participants:	
Descriptions annexes:	
Remarques:	Témoin s'exprime aisément ; témoignage très riche sur son travail
Fiche modifiée par:	Ikanaan
Date d'enregistrement:	01/04/1969
Droits:	Copyright (c) 2012 Université d'Orléans/LLLFreely available for non-commercial use. This file is licensed under a Creative Commons License.
Format:	(IANA MIME Media Type: audio/x-wav)
Durée:	01:00:00
Acoustique:	Excellente
Précisions acoustiques:	très bonne, témoin parle clairement
Lieu spatial:	Orléans
Lieu TGN:	7008337
Lieu Point:	east=1.904; north=47.902
Locuteurs:	• 001LOC1 • 001LOC2 • BA725 • OU
Transcriptions:	• ESLO1_ENT_001_A • ESLO1_ENT_001_B • ESLO1_ENT_001_C



Fiche locuteur

Identifiant locuteur : BA725

Anonyme:	OUI
Année de naissance:	1912
Tranche d'âge:	55/65
Lieu de naissance:	loiret
Sexe:	Homme
Niveau d'études:	CEP
Commentaire:	Enseignement primaire à Orléans
Age de fin d'études:	14
Catégorie Professionnelle (INSEE):	Artisans, commerçants et chefs d'entreprise
Profession en termes propres:	boucher, gérant boucherie supermarché
Langue(s):	Français
Commentaire niveau langue:	
Situation de famille:	Marié
Année d'arrivée:	1915
Domicile:	Orléans centre
Nombre d'enfants:	2
Information sur les enfants:	fils 1 : coiffeur, fils 2 ?
Remarques diverses:	Famille : femme sans activité
Fiche modifiée par:	cdugua
Enregistrements et transcriptions:	• Enregistrement ESLO1_ENTCONT_201 • Transcription ESLO1_ENTCONT_201_A • Transcription ESLO1_ENTCONT_201_B • Transcription ESLO1_ENTCONT_201_C • Enregistrement ESLO1_REPAS_270 • Transcription ESLO1_REPAS_270_A • Transcription ESLO1_REPAS_270_B • Transcription ESLO1_REPAS_270_C • Enregistrement ESLO1_VISIT_601 • Transcription ESLO1_VISIT_601_A • Transcription ESLO1_VISIT_601_B • Transcription ESLO1_VISIT_601_C • Enregistrement ESLO1_VISIT_602 • Transcription ESLO1_VISIT_602_A • Transcription ESLO1_VISIT_602_B • Transcription ESLO1_VISIT_602_C • Enregistrement ESLO1_ENT_001 • Transcription ESLO1_ENT_001_A • Transcription ESLO1_ENT_001_B • Transcription ESLO1_ENT_001_C • Enregistrement ESLO1_ENTCONT_202 • Transcription ESLO1_ENTCONT_202_A • Transcription ESLO1_ENTCONT_202_B • Transcription ESLO1_ENTCONT_202_C • Enregistrement ESLO1_ENTCONT_203

soit par l'entrepôt COCOON *[corpus de la parole]* :

coCOon

COllections de COrpus Oraux Numériques

Accueil | Présentation | Accès aux corpus | Documentation

Accueil > Chercher une ressource > Métadonnées

[fr] ESLO2: entretien 1001

Laboratoire Ligérien de Linguistique

Laboratoire Ligérien de Linguistique (depositor) ; Baude, Olivier (researcher) ; BV1 (speaker) ; BV1AMI (speaker) ; ch_OB1 (speaker)

(création: 2010-01-20; mise à disposition: 2014-07-04; archivage: 2014-12-04T21:57:41+01:00; dernière modification de la notice: 2015-06-12)

Editeur(s):	Laboratoire Ligérien de Linguistique
Description(s):	[fr] Extrait de la seconde Enquête Sociolinguistique à Orléans réalisée autour des années 2010.
Table(s) des matières:	[fr] Trame questionnaire ESLO2-Janvier 2010
Source(s):	[fr] Bande magnétique 1001
Type(s):	Type(s) linguistique: primary_text Type(s) de discours: dialogue Enregistrement sonore
Sujet(s):	Champ(s) linguistique: text_and_corpus_linguistics [fr] Français (code ISO-639: fra)
Langue(s):	[fr] Français (code ISO-639: fra)
Format(s):	(IANA MIME Media Type: audio/x-wav) durée: 0:1H8:43
Droits:	Copyright (c) 2012 Université d'Orléans/LLL Freely available for non-commercial use
Pour citer la ressource:	http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO2_ENT_1001 ou ark:/87895/1.17-475022 ou hdl:10670/1.030e0u

Enfin, certaines métadonnées sont contenues dans le fichier XML de transcription. La racine du document est l'élément <Trans> qui contient les informations suivantes : le nom du transcripteur, le numéro de l'enregistrement et la version, et la date de la transcription :

<Trans scribe="Panot" audio_filename="001" version="15" version_date="091210">

Ensuite vient l'élément <Topics> qui décrit des thématiques de l'interview :

<Topics>

<Topic id="to1" desc="QP1"/>

<Topic id="to2" desc="QP2"/>

Archivage *[Corpus de la parole]*

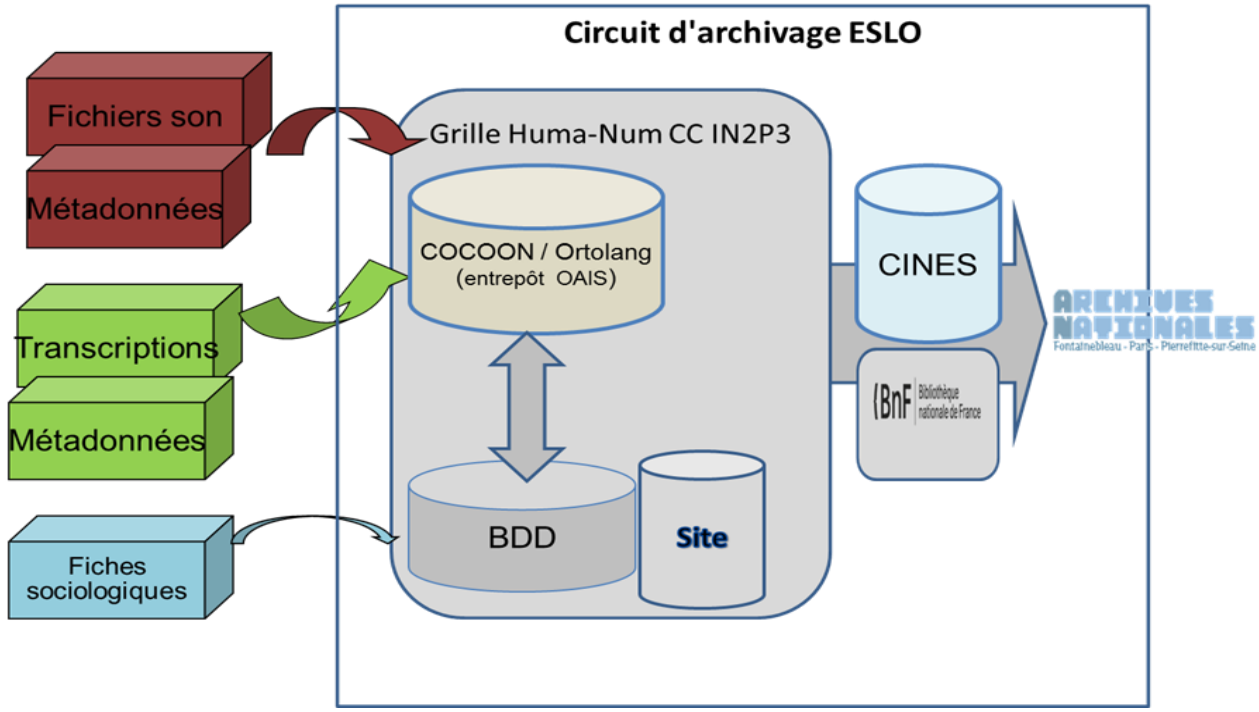
L'application ESLO est une application de gestion pour les enquêtes ESLOs. La gestion couvre les fonctionnalités d'ajout de nouvelles données (principalement enregistrements et transcriptions) de description de celles-ci, de mise à jour des données et métadonnées et de consultation. Suivant les profils des utilisateurs de l'application, ces fonctionnalités peuvent se réaliser différemment.

Une nouvelle fonctionnalité d'archivage a été ajoutée à l'application ESLO afin de piloter les échanges entre celle-ci et l'application Cocoon. Cette fonctionnalité n'est accessible qu'aux administrateurs et leur permet de déclencher des actions liées au cycle de vie des documents.

Un document nait avec le statut « en cours d'édition » et les utilisateurs ayant le profil éditeur peuvent en modifier la description. Un administrateur peut souhaiter figer l'état d'un document et demander son transfert à l'application Cocoon qui se chargera de l'archiver et de le rendre accessible. Une fois la demande de transfert effectuée, le document acquiert alors un statut « en cours d'archivage » et toutes les fonctions d'édition sont bloquées pour les éditeurs. La demande de transfert déclenche du côté de l'application ESLO le dépôt dans un répertoire de transfert d'un paquet d'information contenant : le document à archiver (transcription ou enregistrement) et un fichier de métadonnées le décrivant. Par un traitement en lots, l'application Cocoon va parcourir l'ensemble des paquets déposés, contrôler leur bon formatage et, si aucune anomalie n'est relevée, copier ces informations dans son propre système. Dans tous les cas, l'application après le parcours du répertoire de partage dépose un bordereau signalant les éléments pris en charge et ceux comportant une anomalie. A la vue de ce bordereau, l'application ESLO va, pour chaque document listé, modifier son statut : soit avec la valeur « archivé », soit, en cas d'anomalie, revenir à la valeur « en cours d'archivage » en signalant la raison de l'anomalie.

Ce module de gestion des communications entre ESLO et Cocoon permet également d'envoyer à Cocoon les métadonnées modifiées de documents déjà archivés ainsi que d'envoyer une nouvelle version d'un document déjà archivé.

L'ensemble de ces opérations permet le process suivant qui assure un archivage du corpus ESLOs dans le respect des bonnes pratiques :



Module d'accès à l'application WEB ESLO

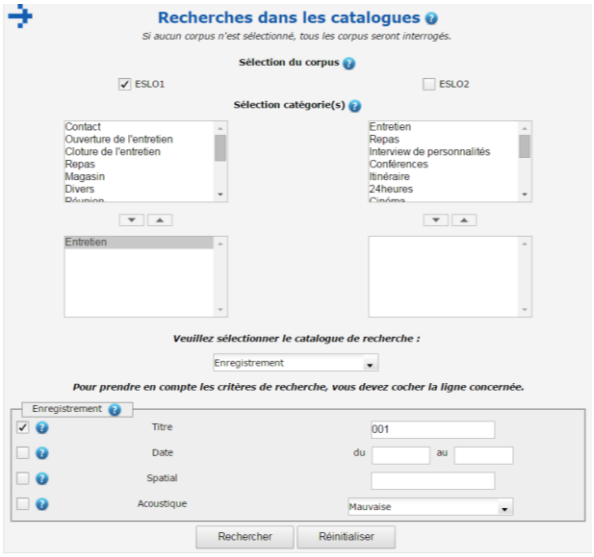
La gestion du corpus ESLO repose sur le modèle OAIS ce qui permet de dissocier la gestion du corpus des interfaces d'accès qui peuvent être multiples. L'application ESLO offre un accès au corpus à partir d'un objectif particulièrement apprécié par les auteurs du projet : permettre une navigation dans le corpus à partir de requêtes sur les formes linguistiques et sur les données sociologiques des locuteurs à partir de données situées.

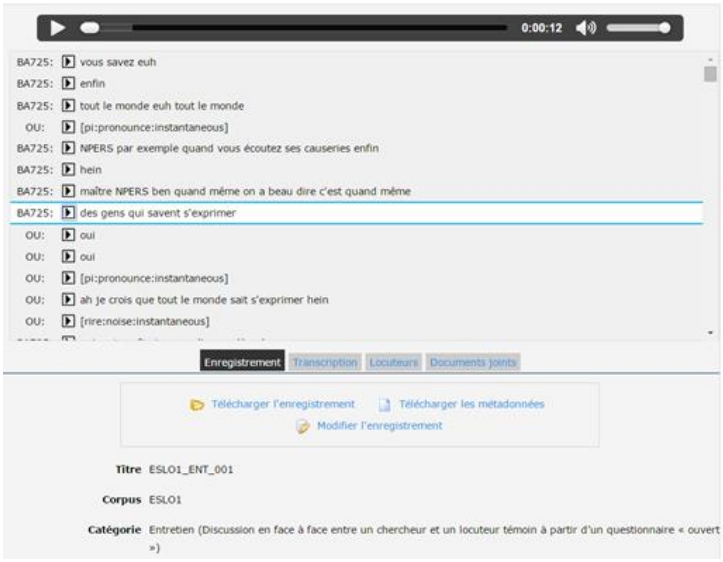
Enfin le site d'accès est prévu pour une totale mise à disposition des données, métadonnées et documents contextuels qui peuvent tous être téléchargés.

L'interface présente donc un accès aux documents sur la méthodologie de l'enquête, les objectifs scientifiques et les documents contextuels :

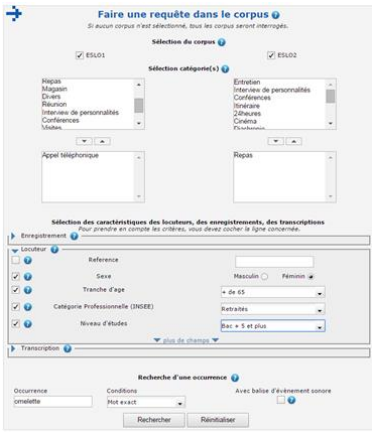
L'interface permet ensuite un accès aux données sous plusieurs formes :

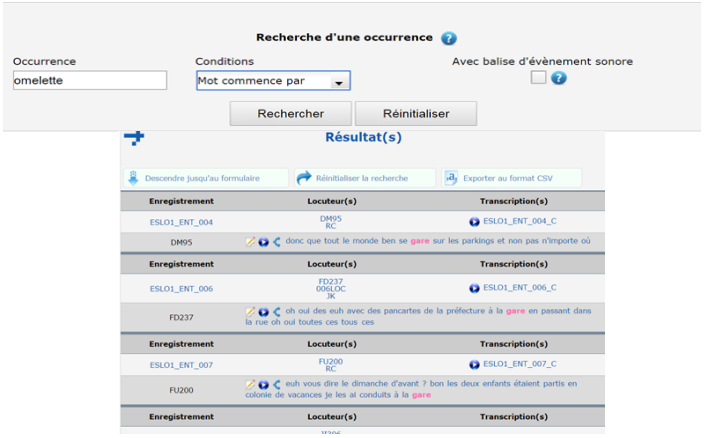
- Le feuilletage d'un catalogue qui donne accès aux enregistrements et aux transcriptions dans leur intégralité. Dans ce cas le croisement avec les données sociologiques et les métadonnées est possible :





- Des recherches sur une forme à partir d'une chaine de caractères :





- L'exploration du lexique :

Résultat(s) d'exploration du lexique

Descendre jusqu'au formulaire | Réinitialiser la recherche

Rang	Fréquence	Mot
1	128688	euh
2	94342	oui
3	93681	de
4	74436	c'est
5	68291	que
6	66825	et
7	64731	hm
8	64028	je
9	58003	pas
10	55110	à
11	54927	vous
12	53058	ça
13	52636	le
14	52599	la
15	50969	les
16	46335	?
17	45411	a
18	45044	des
19	41950	on
20	41250	un
21	38610	qui
22	37402	en

Présentons au passage, pour l'anecdote, la comparaison du lexique entre ESLO1 et ESLO2 pour les 20 formes les plus fréquentes :

ESLO1

Rang	Fréquence	Mot
1	82827	oui
2	72784	de
3	71306	euh
4	55807	que
5	51292	c'est
6	50098	vous
7	47794	je
8	47226	et
9	44696	pas
10	43358	à
11	41524	les
12	39695	la
13	39141	le
14	38581	ça
15	35717	?
16	34927	des
17	31123	a
18	30753	qui
19	30478	un
20	28470	on

ESLO2

Rang	Fréquence	Mot
1	139400	euh
2	115719	de
3	109296	oui
4	90380	c'est
5	84111	que
6	81008	et
7	77441	je
8	71671	pas
9	67932	vous
10	67748	à
11	65789	la
12	64662	ça
13	64570	le
14	63660	les
15	62247	hm
16	55531	des
17	54340	a
18	51079	on
19	50411	un
20	47876	?

3.6.3 Aspects juridiques *[GBP]* *[retour]*

Dans un chapitre précédent nous avons présenté les aspects juridiques et éthiques liés à la phase de collecte. Ceux-ci anticipaient les questions posées par la gestion d'un corpus numérique dans un but de diffusion, de réutilisation et d'archivage.

Protection des données personnelles et anonymisation *[GBP]*

En complément de la démarche de collecte du consentement éclairé, la gestion des données personnelles donne lieu à deux opérations lors de la gestion du corpus : le codage des locuteurs et l'anonymisation des données au sein des enregistrements et des transcriptions.

L'anonymisation actuelle dans ESLO est semi-automatique et porte sur les enregistrements, les transcriptions et les métadonnées. Dans la chaine de traitement du corpus, la phase d'anonymisation est fractionnée ; elle précède la phase de transcription, coïncide avec elle et la suit. Un guide a été rédigé qui décrit la procédure :

(https://www.nakala.fr/nakala/data/11280/9ce4049c)

La première étape d'anonymisation concerne les métadonnées. Elle repose sur le codage des noms propres, l'extraction de l'adresse (ces deux informations étant conservées mais non disponibles) et la correction des données de géolocalisation.

Le codage des noms propres des locuteurs est l'action la plus classique et attendue dans une procédure d'anonymisation. Deux types de codages sont mis en place : les codes aléatoires qui sont générés par notre application suite à la création d'une fiche en saisissant les métadonnées du locuteur (ex : DC738, Fiche du *locuteur*), et les codes répondant à un plan de nommage (ex : DC738FEM).

Fiche du locuteur

Fiche locuteur

Identifiant locuteur : DC738

Anonyme:	OUI
Année de naissance:	1927
Tranche d'âge:	35/45
Lieu de naissance:	Non renseigné
Sexe:	Homme
Niveau d'études:	CEP
Commentaire:	
Age de fin d'études:	16
Catégorie Professionnelle (INSEE):	Artisans, commerçants et chefs d'entreprise
Profession en termes propres:	coiffeur
Langue(s):	
Commentaire niveau langue:	
Situation de famille:	Marié
Année d'arrivée:	1959
Domicile:	Orléans
Nombre d'enfants:	2
Information sur les enfants:	fille, coiffeuse BEPC fils, 12 ans, CES
Remarques diverses:	femme secrétaire Enseignement : cours complémentaires à Beaugency diplôme: CEP (niveau du brevet)
Fiche modifiée par:	pphilardeau
Enregistrements et transcriptions:	• Enregistrement ESLO1_ENT_057 • Transcription ESLO1_ENT_057_A • Transcription ESLO1_ENT_057_B • Transcription ESLO1_ENT_057_C

Le plan de nommage ESLO propose des combinaisons permettant de marquer des relations ou des catégories. En effet, pour marquer les relations, certains locuteurs possèdent des codes construits sur le code aléatoire d'un autre locuteur (BA725FIL). Un autre type de codes du plan de nommage repose sur les numéros des enregistrements, et permet de marquer une catégorie (653CLI, 653VEN, 308STAN dans un appel téléphonique). De plus, les locuteurs non identifiés, non attendus dans un enregistrement et pour lesquels aucune information n'est repérable ni fournie sont aussi codés en lien avec l'enregistrement (452INC).

Toujours au niveau des métadonnées, nous intervenons au niveau du lieu de l'enregistrement. L'anonymisation s'effectue à travers la transformation de l'adresse en coordonnées GPS de manière à délimiter un périmètre correspondant au « pâté de maisons ».

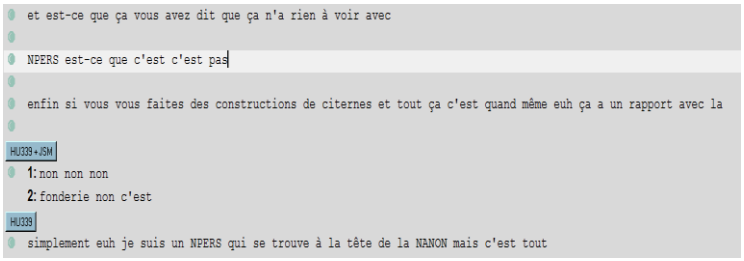
Notons que les données nominatives (nom, adresse, numéro de téléphone) des témoins sont conservées dans une BDD physiquement indépendante, conformément aux recommandations de la CNIL (cf. Baude *et al.* 2006). Ces informations sont recueillies par le chercheur dans ESLO1 et renseignées dans ESLO2 par les locuteurs eux-mêmes qui complètent un « Formulaire témoin ». Pour ESLO2, ce qui n'était pas le cas pour ESLO1, les témoins signent un « Formulaire de consentement » concernant : leur participation au projet scientifique, la conservation (mais non la diffusion) de leurs informations personnelles dans le seul but d'être recontactés, l'utilisation des enregistrements et de leurs transcriptions pour la recherche et pour la diffusion. Il leur est précisé que l'enregistrement et les

transcriptions seront rendues anonymes (nom remplacé par un code et son brouillé pour la séquence concernée). Il s'agit donc là d'un consentement éclairé qui repose sur un document contractuel qui décrit clairement la manière dont les données seront traitées et les différents types d'utilisation dont elles feront l'objet. En ce sens, on peut considérer que les témoins sont informés des « risques » de leur participation au projet.

La conservation des données nominatives s'est révélée particulièrement intéressante dans le cas d'ESLO puisque cela a permis, quarante ans après la première enquête (ESLO1), de retrouver des témoins et de conduire de nouveaux entretiens avec eux. Un module d'ESLO2 qui offre des possibilités riches et intéressantes pour des recherches diachroniques a ainsi été constitué.

La question du codage du locuteur ayant déjà été traitée au niveau des métadonnées, les codes sont repris dans les transcriptions. Il reste alors à traiter les données identifiantes contenues dans les énoncés. La deuxième intervention se situe donc au niveau de la transcription. Il est demandé aux transcripteurs de remplacer par l'hyperonyme NPERS les noms de personnes et par NANON les autres segments du discours permettant d'identifier un locuteur – i.e. le faisceau d'indices identifiants –, ou encore des propos « sensibles ». Ces opérations sont par la suite vérifiées par un chercheur avant qu'elles ne soient traitées au niveau de l'enregistrement, ce qui constitue la troisième étape de l'anonymisation. Il s'agit donc d'une action faite par un humain sans aucun repérage automatique car elle ne représente pas une charge de travail signifiante du fait de son intégration préalable à la phase de transcription.

Anonymisation dans la transcription

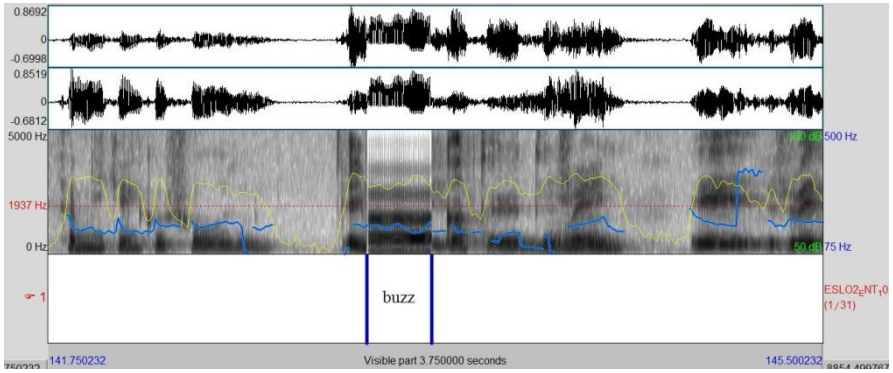


Le fichier de transcription contient néanmoins des données nominatives. L'entête du fichier XML indique en effet le nom de la personne qui a effectué la transcription. Cette information est conservée avec l'accord de l'auteur de la transcription afin de permettre la traçabilité de l'activité scientifique et de respecter la propriété intellectuelle du travail fourni (cette démarche correspond à l'une des recommandations de la charte Ethique et Big Data).

Une dernière étape est effectuée avec le logiciel Praat grâce à un script réalisé par Daniel Hirst (LPL, Aix-en-Provence) qui permet de modifier automatiquement les segments du fichier son à partir d'un repérage dans les annotations. Les segments sonores concernés sont modifiés afin que l'on ne puisse pas comprendre ce qui est prononcé tout en conservant

certaines caractéristiques comme, notamment, la courbe intonative. Après un repérage temporel des NPERS et des NANON dans la transcription – un travail manuel qui est devenu automatique grâce à une application développée par Flora Badin (LLL, Orléans) –, l'isolement des segments concernés est effectuée sous Praat après la création d'un textgrid. Le code « buzz » marqué dans chacun des segments délimités permet au script d'opérer à l'intérieur des balises temporelles et de brouiller le signal.

Anonymisation du son



L'anonymisation et les faisceaux de données identifiantes *[GBP]*

Parallèlement, une équipe du projet ESLO (Iris Eshkol-Taravella, Olivier Baude, Céline Dugua, Denis Maurel) a entrepris une étude sur le traitement automatique de l'anonymisation. La partie TAL est détaillée dans les travaux d'Eshkol-Taravella (HDR, 2015). L'intérêt de ce travail est de partir d'une contrainte juridique : déterminer les données identifiantes dans un corpus oral pour porter une réflexion sur la catégorisation de ce type de données. En effet, cette expérience permet de constater que la notion d'entité nommée n'apporte que peu de réponse et qu'il convient de définir une catégorie représentative d'un « faisceau d'indices identifiants » :

Recherche des indices permettant une identification: l'anonymisation des transcriptions du corpus ESLO (Eshkol et al. 2015)

L'impossibilité d'identifier est une notion complexe qu'on a trop souvent réduite à l'effacement des noms propres. La tâche est bien plus difficile, mais aussi plus stimulante pour les recherches en linguistique et en TAL.

L'anonymisation relève de procédures différentes selon qu'on traite l'enregistrement sonore, sa transcription ou les métadonnées descriptives. Toutefois, dans tous les cas, l'objectif reste le même. Si selon certains juristes la voix est une donnée identifiante ce qui nécessiterait de modifier le signal acoustique de tout enregistrement et par là même obèrerait toute recherche en linguistique, les pratiques des chercheurs s'orientent plus généralement vers un

traitement des données personnelles au sens large. Que ce soit sur l'oral ou sur l'écrit celles-ci sont diverses, il peut s'agir d'une forme nominative, d'une profession, d'un statut, d'une caractéristique physique, etc. et/ou du recoupement de plusieurs de ces informations. Si l'on convient que l'anonymisation ne se réduit pas à l'effacement des noms propres, il est nécessaire de définir avec précision quels sont les traitements à effectuer pour répondre à l'objectif de réduire les possibilités d'identification. Dans le cas de grands corpus, ces traitements deviennent une étape fondamentale du travail de constitution du corpus avec des effets très importants sur la gestion et la diffusion des données.

(…) Si l'on veut anonymiser efficacement le discours, on ne peut pas s'arrêter aux entités nommées car d'autres indices peuvent renvoyer vers le locuteur ou vers la personne dont il parle. Observons les exemples tirés du corpus ESLO1 :

- j'ai une maladie du foie ça m'a même occasionné une petite scoliose déformation légère de la colonne vertébrale.

- mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville

- je suis scout de France le jeudi soir où j'anime un un atelier photos

Cette catégorie des indices est large. Elle inclut des éléments assez hétérogènes désignant les différentes informations personnelles sur la personne : évènements, activités sociales, loisirs, maladies, handicap, etc. qui peuvent au même titre que le travail, la famille donner les informations sur le locuteur ou la personne dont on parle.

Ainsi, le « faisceau d'indices » inclut les entités nommées identifiantes, mais peut contenir aussi d'autres éléments qui permettent l'identification soit directement, soit, par combinaison au sein de ce faisceau : la personne est patron d'un bar au moment d'enregistrement, et avant elle travaillait dans l'aviation militaire. Le processus d'identification est progressif, il se construit au fur et à mesure de l'accroissement des indices. On peut supposer qu'un indice identifiant ou une série de ces indices est associée à un individu particulier dans la mémoire à l'aide d'un certain lien dénominatif qui sera réactivé lors de leur apparition dans le discours. C'est grâce aux facteurs contextuels, c'est-à-dire grâce aux connaissances que l'utilisateur du corpus maitrise concernant le locuteur ou la personne mentionnée dans le discours de celui-ci, que l'identification peut se faire.

Aspects juridiques et propriété intellectuelle *[GBP]*

Le second aspect juridique concerne la propriété intellectuelle des données du corpus. Celles-ci sont considérées comme des productions du laboratoire de recherche dans un but de diffusion libre de l'information scientifique. Il a donc été décidé que toutes les données et tous les documents du corpus des ESLOs doivent être délivrés sous licences Creative Commons http://creativecommons.fr/

Les enregistrements et les transcriptions d'origine relèvent de la licence, car le consentement éclairé stipule qu'il n'y aura pas d'utilisation commerciale sans autorisation :

Attribution + Pas d'Utilisation Commerciale + Partage dans les mêmes conditions (BY NC SA) : 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.

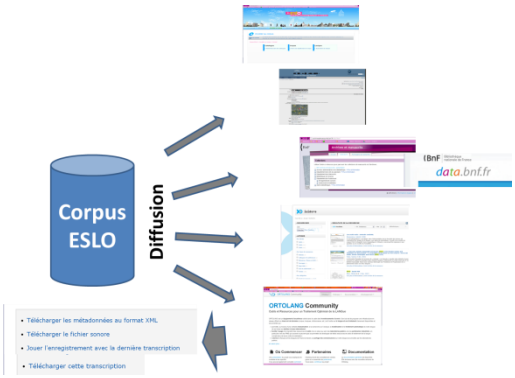
Afin de faciliter la réutilisation des versions annotées, celles-ci sont délivrées sous la licence :

Attribution + Partage dans les mêmes conditions (BY SA) : Le titulaire des droits autorise toute utilisation de l'œuvre originale (y compris à des fins 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.

Toutes ces informations sont indiquées dans les métadonnées.

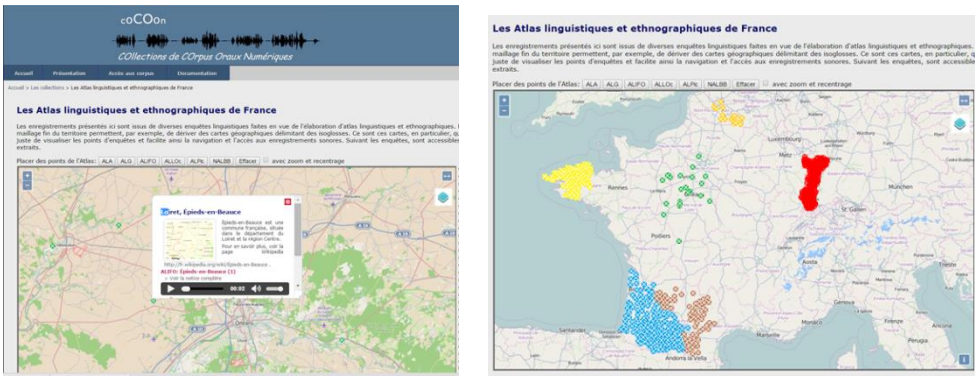
3.6.4 Diffusion du corpus ESLO [retour]

La conservation du corpus ESLO dans un entrepôt OAI et les choix de gestion des droits facilitent la diffusion de celui-ci. En 2015 la diffusion est effectuée directement par différentes plateformes qui apportent toutes leurs spécificités :



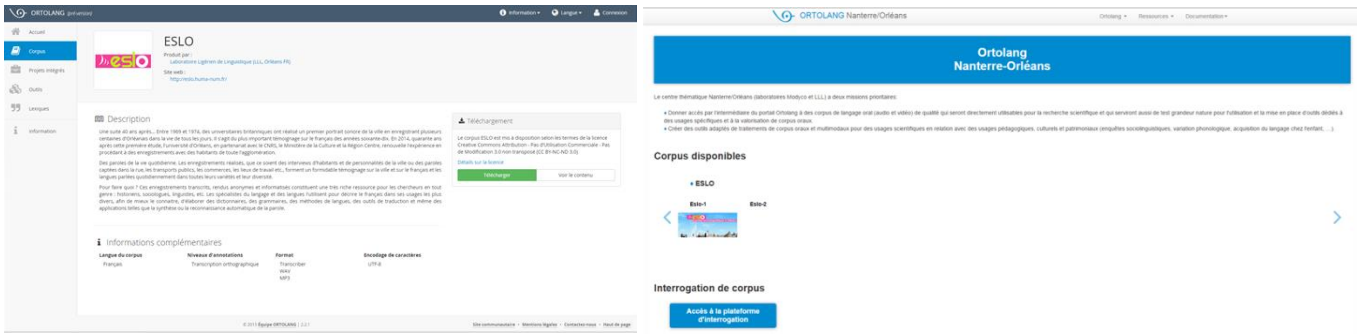
- La plateforme ESLO
  Cf. *infra*.
- La plateforme COCOON [Linked Open Data]
  Outre sa fonctionnalité d'archivage, la plateforme COCOON intègre le corpus ESLO à d'autres collections de corpus oraux. Elle offre d'autres fonctionnalités liées à l'usage du format RDF et au modèle du Linked Data (« Le modèle du Linked Open Data appliqué à des ressources orales » (Jacobson & Baude, soumis)). Cette perspective novatrice est détaillée dans un autre chapitre).

Accès par cartes et usage du *LOD* sur la plateforme Cocoon



- La plateforme Ortolang

  L'objectif de la plateforme Ortolang est de donner un accès unifié aux corpus, ressources et outils linguistiques. Le Corpus ESLO est diffusé au sein de la plateforme et par l'intermédiaire d'un centre dédié aux corpus oraux et multimodaux.



- La BnF *[Viaut-BAM-EAD]*

  Le corpus ESLO n'est pas directement diffusé par la BnF mais il est signalé au sein de ses catalogues et disponible (dans le cas d' ESLO1) pour une consultation sur place.

  Pour ESLO, les supports ainsi que la documentation papier d'origine ont fait l'objet d'un don à la BnF. Celle-ci a catalogué et décrit ces objets d'archives dans un instrument de recherche en EAD et publié dans BAM (Catalogue Archives et manuscrits). Une partie des documents a été numérisée avec l'objectif de préserver l'information des ravages du temps et d'en faciliter l'accès. La consultation des

documents est réservée aux chercheurs dans les locaux de la BnF alors que les métadonnées sont librement accessibles sur la Toile. Ces bandes ont été également numérisées par le laboratoire LLL, qui a procédé à un travail d'anonymisation et de transcription. L'ensemble des données est déposé dans Cocoon afin d'en permettre un accès libre.

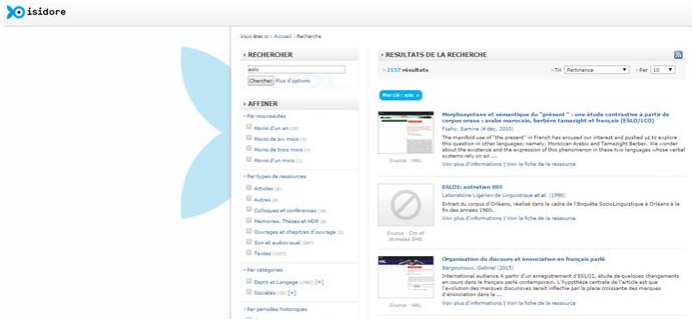
Les deux gisements d'informations BAM et Cocoon ne comportent donc pas les mêmes données mais des données qui ont un lien phylogénétique (les copies numériques et anonymisées de Cocoon proviennent des supports conservés à la BnF). Ces données peuvent aussi être complémentaires : la documentation d'origine est à la BnF, les transcriptions sont dans Cocoon. Enfin, les descriptions (métadonnées) qui ont été faites de part et d'autre partagent de nombreuses informations. Afin que l'utilisateur puisse naviguer aisément dans les deux catalogues, des liens ont été établis dans les deux sens. Dans Cocoon un lien systématique sur chaque ressource a été ajouté pour donner l'URL dans BAM de la description du support. Dans l'autre sens, un lien a été ajouté dans BAM pour renvoyer à Cocoon pour une consultation hors les murs (en version anonymisée) ainsi que pour accéder aux informations complémentaires (les transcriptions).

L'utilisation de BAM permet également une intégration au projet de Web sémantique de la BnF : Data.bnf.fr

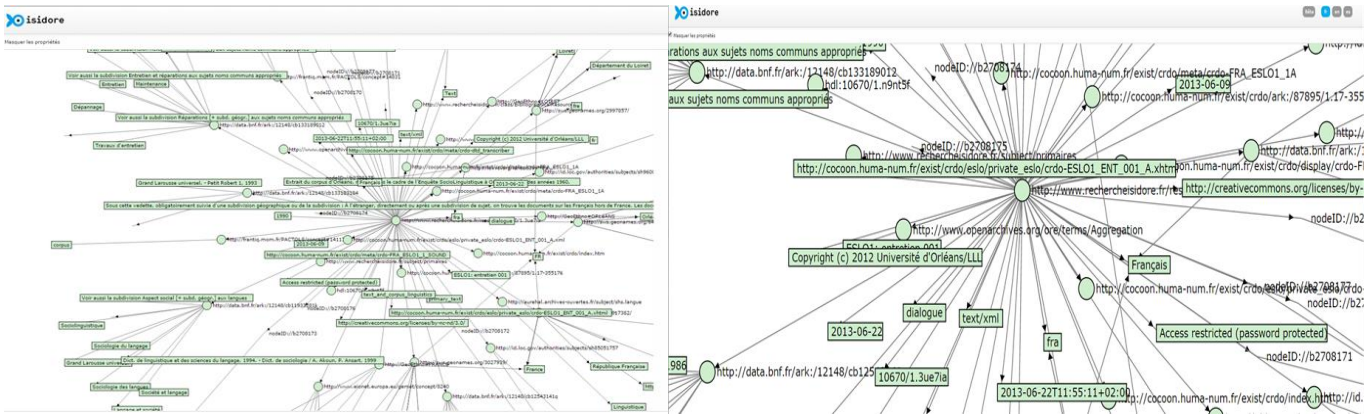
- ISIDORE
  La diffusion d'ESLO par Cocoon permet également un signalement dans ISIDORE :
  « *ISIDORE est un service qui collecte, enrichit et offre un signalement et un accès unifié aux documents et données numériques des sciences humaines et sociales. ISIDORE « moissonne » – c'est le terme consacré – les notices, les métadonnées et le texte intégral issus des publications électroniques, des corpus, des bases de données et des actualités scientifiques, accessibles sur le web et proposés dans des standards ouverts d'interopérabilité. (…) Une fois moissonnées, ces informations sont enrichies en trois langues (anglais, espagnol et français) par croisement avec des référentiels métiers. (…) Les enrichissements multilingues permettent de relier les données entre*

elles. Ces informations constituent des points d'entrée vers le texte intégral qui est lui aussi indexé quand cela est possible. (…) Utilisant les méthodes et principes du web de données (modèle RDF) et du linked data (URIs), (…)ISIDORE associe plus d'une centaine de producteurs de données et moissonne plus de 2000 sources de données : les principales plateformes d'édition électronique, un très grand nombre de bibliothèques (de recherche, universitaires, municipales) mais aussi de nombreuses bases de données des SHS. »

Résultats d'une requête ESLO dans Isidore



Visualisation d'un graphe RDF « ESLO_ Entretien 001 » dans Isidore :



3.7 Un exemple d'analyses : « la liaison dans les ESLOs » [retour]

Articles et livre :	
	○ 2011, (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?, https://hal.archives-ouvertes.fr/hal-01162479 ○ 2015, Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ?, https://halshs.archives-ouvertes.fr/halshs-01163047 ○ (à paraitre) Jean Zay et la mémoire orale : du politique au scientifique, de l'État à la ville, d'hier à aujourd'hui), https://halshs.archives-ouvertes.fr/halshs-01165944
Communications orales :	
	○ 2011, La variation en réserve, https://halshs.archives-ouvertes.fr/halshs-01165946 ○ 2012, Paroles élues, la liaison dans les discours des hommes politiques, https://halshs.archives-ouvertes.fr/halshs-01165938 ○ 2013, Regards croisés sur la liaison dans le corpus ESLO, https://halshs.archives-ouvertes.fr/halshs-01165936
Documents :	
	○

Le projet ESLO a été conçu autour d'une rupture de linéarité dans la démarche qui partirait de la collecte des données vers l'analyse au profit d'une interaction systématique entre les différentes étapes, niveaux, procédures méthodologiques et théoriques.

L'étude de la liaison en français a accompagné cette démarche pour plusieurs raisons. Depuis les travaux d'Encrevé (Encrevé 1983[100], 1988[101]), la liaison a été au cœur d'avancées fortes en sociolinguistique. Ainsi, le projet PFC (Durand et al., 2011[102]) , notamment dédié à la liaison, offre une base de comparaison d'un très grand corpus qui contient des dizaines de points d'enquête en France. Le corpus des ESLO a donné lieu à une thèse fameuse sur la sociophonologie de la liaison orléanaise (De Jong 1994[103]). Par ailleurs ce sont bien des phonologues que provient la réflexion la plus poussée sur le rôle de la linguistique de corpus en linguistique (Laks 2008[104], 2010[105], Scheer 2004[106]).

3.7.1 Approches théoriques de la liaison [retour]

De nombreux travaux s'attachent au phénomène de liaison en français. Ils se retrouvent dans divers domaines de recherche:

Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ? (Baude & Dugua 2015:350)

[100] ENCREVÉ, P. (1983). « La liaison sans enchainement ».
[101] ENCREVÉ, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*.
[102] DURAND, J. et al. (2011). « Que savons-nous de la liaison aujourd'hui? ».
[103] DE JONG, D. (1994). « La sociophonologie de la liaison orléanaise ».
[104] LAKS, B. (2008). « Pour une phonologie de corpus ».
[105] LAKS, B. (2010). « Langage et variation : pourquoi y a-t-il de la variation plutôt que rien ? ».
[106] SCHEER, T. (2004). « Le corpus heuristique : un outil qui montre mais ne démontre pas ».

La liaison est un processus généralement décrit comme principalement phonologique ; elle a d'ailleurs très souvent servi de phénomène test pour l'expression des représentations phonologiques (Schane 1967 ; Encrevé 1988 ; Tranel 1996 ; Côté 2005). Son évolution diachronique en lien avec les évolutions de l'orthographe trouve place dans les travaux de linguistique historique (Clédat 1917 ; Fouché 1952 ; Bourciez & Bourciez 1971). Au-delà des traitements purement phonologiques, des auteurs ont montré, par exemple, que la liaison en /z/ peut être traitée comme un marqueur morphologique du pluriel (Morin et Kaye 1982 ; Morin 2003 [1998]). Des critères de groupes syntaxiques et sémantiques (Laks 2005), comme des critères de groupes prosodiques (Grammont 1914 ; De Jong 1990) sont utilisés pour délimiter des contextes de liaisons possibles. Les chercheurs dans le domaine du traitement cognitif du langage s'appuient également sur la liaison pour cerner des phénomènes qui lui sont liés, tels la segmentation du lexique et le traitement on-line de l'accès lexical (Spinelli, Cutler & McQueen 2002 ; Spinelli & Meunier 2005). Depuis les années 2000, les recherches en acquisition s'efforcent de décrire et formaliser les étapes de mise en place de ce phénomène chez les jeunes enfants en langue première (Chevrot, Dugua & Fayol 2009 ; Nardy & Dugua 2011 ; Wauquier & Braud 2005) et en langue seconde (Harnois-Delpiano et al. 2012 ; Wauquier 2009). Certains travaux de sociolinguistes (Encrevé 1988 ; Gadet 2003) utilisent la liaison en tant que phénomène de variation pour préciser, par exemple, les facteurs extralinguistiques influant sur la réalisation ou la non réalisation de cette unité phonologique variable.

On peut résumer les grandes approches de la phonologie en quatre domaines comme le fait Dugua (2006[107]) :

- Approche à partir de règles (ex. : Schane) qui permettent de passer de formes sous-jacentes à des formes de surface, e.g. des règles qui suppriment des consonnes dans certains contextes (la consonne de liaisons (CL) n'est pas prononcée devant un mot commençant par une consonne = règle de troncation).

- Approche multilinéaire (ex. : Encrevé, Wauquier). La représentation des formes phonologiques sous-jacentes, en intégrant différents niveaux phonologiques (squelette, syllabe, phonème), substitue aux règles des principes généraux et des paramètres différents en fonction des langues. Ici la CL est considérée comme « flottante ».

- Théorie de l'optimalité (ex. : Tranel). Proche/issue des modèles connexionnistes, elle fait intervenir le système cognitif qui, en phonologie, fonctionne selon des contraintes multiples qui doivent être satisfaites. Ce ne sont plus les représentations (sous-jacentes) qui importent mais la hiérarchisation des contraintes qui détermine la réalisation phonétique.

[107] DUGUA, C. (2006). *Liaison, segmentation lexicale et schémas syntaxiques entre 2 et 6 ans. Un modèle développemental basé sur l'usage*.

- Théories basées sur l'usage (et grammaires de constructions) (ex. : Bybee). Ni règles ni contraintes : les formes lexicales sont stockées sur la base des formes entendues et mémorisées d'où émergent des schémas abstraits qui permettront de générer tout type de construction ou d'énoncé.

Comme le souligne B. Laks, la solution à l'un des derniers problèmes majeurs de la phonologie nécessite une étude multidimensionnelle pour laquelle l'utilisation, dans une perspective variationniste, de grands corpus peut être un élément essentiel :

> *« … la liaison est un indicateur social explicite, un des rares lieux de la langue où les plus anti-variationnistes des linguistes ont été amenés à reconnaitre la variation sociale et l'hétérogénéité linguistique »* (Encrevé, 1983 :42-43)

Dans cette perspective le projet PFC plaide pour la réalisation d'un grand corpus dédié à l'analyse de la liaison (Durand & laks 2011:110-111):

> *« S'il est une dimension du phénomène de la liaison en français qui a retenu l'attention c'est bien son extrême variabilité socio-stylistique, au moins au sein de données captées à la volée. Pourtant, on serait bien en peine de citer une enquête de grande envergure permettant d'illustrer et de quantifier les différents aspects de cette variation. Si les variations diatopiques, diaphasiques, diastratiques et diachroniques de la liaison ont fait l'objet d'un nombre incalculable de commentaires, tous en général repris les uns des autres, la documentation du phénomène appuyée sur des données variationnelles précises et quantitativement pertinentes fait encore largement défaut. P. Encrevé (1988 : 44-45), très sensible on le sait aux dimensions variationnelles, déplore l'absence d'enquêtes et de données fiables. Recensant l'existant et le disponible, il souligne la pauvreté de la documentation empirique. Ne travaillant lui-même que sur un corpus réduit de 21 hommes politiques saisis dans un seul style, il analyse 5 787 contextes. Depuis vingt ans, peu de données d'ampleur se sont ajoutées à celles recensées par P. Encrevé : D. de Jong (1994) a présenté une analyse secondaire du corpus d'Orléans (Blanc & Biggs 1971) portant sur 45 locuteurs et 16 000 contextes ; B. Laks (2007) a publié une analyse portant sur 43 hommes politiques (73 extraits de discours) enregistrés entre 1908 et 1998 avec 2 879 contextes de liaison (cf. également Green & Hintze (1990, 2001) et l'étude microsociologique récente de Ranson 2008). Au total, on constate que les données empiriques restent très parcellaires et que la possibilité d'étudier les différentes dimensions de la variation sont limitées*

par l'unicité de style et par l'absence de variation sociale, géographique ou d'âge. »

L'analyse à partir des données nécessite une répartition des liaisons en liaison *obligatoire-invariable*, liaison *interdite-erratique*, ou *facultatives-variables*, qui est loin d'être une opération neutre (Dugua 2006 :30-37) :

- soit à partir de corpus dans lesquels sont relevés les réalisations des liaisons en fonction des contextes (ex : De Jong),
- Soit à partir d'une approche plus théorique (Sur ce point, voir Dugua 2006 :31).
    - En fonction des catégories syntaxiques (Delattre) : la liaison se fera entre deux catégories syntaxiques dont l'union est forte (pas de pause possible).
    - Traitement fonctionnel : La réalisation ou non de la liaison changerait le sens de l'énoncé : *elle donne un bal* / *elles donnent Tun bal*.
    - Généralisations syntaxiques formelles : prise en compte de la structure de la phrase dans son ensemble.
    - Approche lexicale (De Jong) : considérant que les liaisons sont multiples, il faut regarder, pour chaque forme lexicale, le résultat (ex : liaisons après *soient* vs *sont*).

Notons que, d'une manière générale, théoriser la variation linguistique constitue une difficulté certaine, comme le soulignent Durand et Laks.

« La difficulté à laquelle est confronté Grammont est en fait très récurrente dans l'histoire de la phonologie : s'il existe des lois ou des contraintes, leurs dynamiques sont souvent profondément contradictoires. Les langues sont des systèmes intrinsèquement variables [...]. La question du traitement de l'hétérogénéité et la nécessité de construire des modèles théoriques et formels intégrant totalement la variation inhérente restent d'une très vive actualité en phonologie moderne (Durand & Laks, 2000: 36-37). »

Outre la classification, la question du statut des consonnes de liaisons se pose. Trois conceptions sont en concurrence : (i) la CL est rattachée au Mot1 (position la plus classique, en accord avec l'orthographe), (ii) indépendance de la CL (elle est alors épenthétique), (iii) la CL est rattachée au Mot2. A ces conceptions s'ajoutent des conceptions alternatives (3 exemples : Morin & Kaye – Côté – Laks) qui partagent l'idée que la liaison n'est pas un phénomène unique et que par conséquent il n'existe pas un seul et unique statut de la CL à envisager, mais que selon les types de liaison, les contextes, etc. on pourrait avoir des statuts différents. Enfin l'approche des grammaires de construction apportent une nouvelle appréhension du phénomène.

Il reste à déterminer quels sont les facteurs qui déterminent la réalisation ou non des liaisons facultatives. Ceux-ci sont répertoriés par Nardy (Nardy 2008 :107-115[108])

- Facteurs intralinguistiques (p.107)
  - Catégorie grammaticale (catégorie grammaticale du mot1 : les avis divergent)
  - Longueur du mot (plus de liaisons après mot1 court – plus de liaisons devant mot2 mono et bi-syllabique que trisyllabique)
  - Lexique (forme lexicale des mots : ex. *sont* vs *soit*)
  - Nature de la CL (z / t) Dans PFC : z, n, t, r, p (dans l'ordre décroissant de fréquence Durand et al. 2011, p.124)
  - Nature du segment précédent (si le Mot1 se termine par une consonne ou une voyelle)
  - Prosodie
  - Fréquences des mots (des mots1 surtout, mais aussi des collocations mot1-mot2)
- Facteurs extralinguistiques (p.112)
  - Milieu social (données de De Jong)
  - Genre (femme > homme) : De Jong, Malécot. Peu exploité chez Ashby (corpus de Tours)
  - Situation de communication (formel / informel : Agren, Ahmad, Lucci, Encrevé, Moisset)
  - Age

Ces différents éléments théoriques plaident pour une analyse du phénomène de la liaison à partir de corpus variationnistes dont les conditions de production permettent une étude *de données situées*.

3.7.2. La liaison chez Encrevé, les corpus et la sociolinguistique inversée [retour]

L'apport d'Encrevé sur la liaison n'est plus à présenter. Nous nous contenterons de quelques éléments nécessaires pour la suite de cette étude.

Définition de la liaison [Encrevé 1988:23] :

> *"Phénomène ayant lieu dans la chaine parlée au contact entre deux mots, dont le premier lorsqu'il est prononcé isolé ou devant un mot commençant par une consonne (C) se termine par une voyelle (V) et dont le second prononcé isolément commence par une voyelle.*

[108] DUGUA, C. (2008). « « un ours » / « des nours » ou le rôle de la fréquence sur l'acquisition de la liaison en français ».

ex : un peti**T**enfant

Ce phénomène est caractérisé par deux traits phonétiques :

- ***la présence d'une consonne dite consonne de liaison*** *(CL) qui n'apparait que dans ce contexte* [sinon c'est un enchainement]
- ***la ressyllabation*** *qui fait entendre la CL à l'attaque de la première syllabe du second mot en jeu.*

  *Un petiT/ enfant VS Un peti / Tenfant »

Deux points sont importants : d'une part une consonne se fait entendre alors qu'elle est muette dans tous les autres contextes et d'autre part cette consonne se déplace de la fin du premier mot à l'initial du second.

La liaison se décrit par des règles phonologiques :

- ***Phonologie : pur mécanisme phonologique d'enchainement CV***
  (évitement du hiatus)[Schane 1967] ; puis phonologie tridimensionnelle [Encrevé 1988]

LIAIS (obl) : [-syll] # [+syll] = 2
1 2 3 2 1 3

Forme sous-jacente	/((tRop=)$_{A4v}$(fasil=)$_{A4j}$)$_{SA4j}$	/((tRop=)$_{A4v}$(eze=)$_{A4j}$)$_{SA4j}$
Effacement des parenthèses internes	(tRop=fasil=)$_{SA4j}$	(tRop=eze=)$_{SA4j}$
Règle de Troncation	(tRo=fasil=)$_{SA4j}$	La règle ne s'applique pas
Effacement final des parenthèses	tRo=fasil=	tRop=eze=

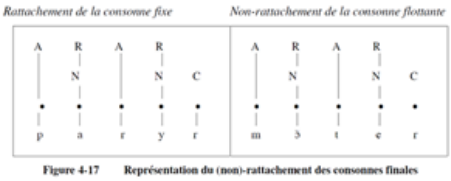


Figure 4-17 Représentation du (non)-rattachement des consonnes finales

Mais sa nature sociale est évidente :

« *La liaison, au contraire, constitue depuis le 17^{e} siècle au moins un enjeu de lutte et de concurrence pour le bon usage (20)* » (Encrevé 1983:42).

La classification classique (en liaisons obligatoires, interdites, facultatives) est revue et chaque élément du classement nécessite une réflexion sur la pertinence de celui-ci . Encrevé reprend la classification des liaisons en variables, invariables et erratiques :

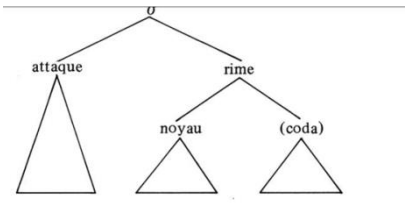
Tableau de classification sommaire des liaisons

	Invariables	Variables	Erratiques
NOM	déterminatif + pronom adjectif nom *vos enfants* *deux autres* *un ancien ami*	nom pluriel + *des soldats anglais* *ses plans ont réussi*	nom singulier + *un soldat anglais* *son plan a réussi*
VERBE	pronom personnel + verbe *ils ont compris* *nous en avons* verbe + pronom personnel *ont-ils compris* *allons-y*	verbe + *je vais essayer* *j'avais entendu dire* *vous êtes invité* *il commençait à lire*	
INVARIABLES		invariables monosyllabiques + *en une journée* *très intéressant* invariables polysyllabiques + *pendant un jour* *toujours utile*	et + *et on l'a fait*
SPÉCIALES	formes figées *comment allez-vous* *les États-Unis* *accent aigu* *tout à coup* *de temps en temps*		h aspiré *des héros* *en haut* + un, huit, onze et dérivés *la cent huitième* *en onze jours*

Delattre, 1966.

Par la suite, Pierre Encrevé constate l'existence de liaisons sans enchainement (LSE), qu'il décrit comme posant un problème sociolinguistique inversé puisque ce sont les locuteurs les plus légitimes qui s'écartent du bon usage :

« *C'est pourquoi il parait utile d'étudier précisément ce qu'il en est de cette forme de liaison non-conforme au bon usage explicite que constitue la liaison non-enchainée, qui est aussi directement liée à un aspect essentiel de la recherche en théorie phonologique aujourd'hui, la théorie prosodique.* » (Encrevé 1983 :43)



« *Le sommet de syllabe, σ, domine les constituants attaque et rime. Une attaque (A) peut comporter une ou plusieurs consonnes, selon les langues ; la rime (R) comprend toujours un noyau comportant au moins une voyelle, et peut comprendre une coda comportant une ou plusieurs consonnes (cf. M. Halle and J. R. Vergnaud, Metrical Phonology, MIT, 1978, mimeo ; J. Kaye et J. Lowenstamm, De la syllabicité, 1981, mimeo)* ».

Les études empiriques de corpus démontrent que la LSE est un phénomène de variation complexe. Ceci n'empêche pas Encrevé de proposer une théorie qui prend en compte cette

variation tout en décrivant des règles : le double flottement. Cette théorie est toujours au cœur des débats sur la liaison :

> *« La liaison dans sa variante non-enchaînée ne manifeste donc pas pour Pierre Encrevé une anomalie marginale, mais elle illustre le fait que la CL est le locus variationis par excellence. Il en propose dans le cadre de la phonologie autosegmentale la représentation formalisée sur la base de l'hypothèse du double flottement.*
> *Cette conception suppose*
> *i) un flottement de la CL sur la ligne segmentale, donc l'absence de rattachement à une position segmentale*
> *ii) un flottement de la CL sur la ligne syllabique, donc l'absence de rattachement à une attaque ou à une coda*
> *iii) une position squelettale disponible permettant l'ancrage de la CL à la fois au plan syllabique (donc en attaque ou en coda) et au plan segmental*
> *iv) la réalisation de CL n'est pas la conséquence d'une dérivation par règles mais le résultat de conventions de bonne formation telles qu'elles ont été paramétrisées pour le français. »* (Wauquier, à paraitre)[109]

Non liaison :

Liaison enchainée :

Liaison sans enchainement :

Cette théorie établie sur une base phonologique dépasse ce domaine pour atteindre « la langue » sans refouler ses aspects sociologiques et cognitifs :

[109] Wauquier S. (à paraître), Des ornithorynques et consonnes « doublement flottantes », pour une théorisation unifiée de la liaison, *Faire signe*, pour Pierre Encrevé.

> *« Pour nous le squelette de positions pures correspond aux nombres de places potentielles définissant un mot donné pour un locuteur donné (car il ne va pas de soi que tous les locuteurs partagent des représentations lexicales absolument identiques) : nous entendons par là le nombre d'unités possibles que le locuteur attribue intuitivement à un mot mémorisé. Nous pensons que cette intuition existe, que les langues soient écrites ou non. Nous suivons ici l'argumentation de Sapir (1933) sur la « réalité psychologique des phonèmes »*. Encrevé (1988 :153)[110]

Dans le cadre du travail présenté ici, c'est bien la question de l'empirisme en linguistique qui se pose prioritairement. Ainsi, Encrevé souligne « le problème des données » :

« Etablir la réalité phonétique de la liaison sans enchainement ne justifie pas de lui faire une place dans la grammaire : de traiter ce fait comme linguistiquement significatif alors qu'il n'a jamais été considéré comme tel. L'absence d'études linguistiques sur ce phénomène peut s'expliquer à la fois par la nature et par la date des données qui sont à la base des travaux proposés aujourd'hui. Fouché et Delattre, qui fournissent (avec Grammont et Martinon qui les inspirent et Léon et Grevisse qui dépendent d'eux) les sources principales des travaux modernes, visent l'un «la conversation soignée des Parisiens cultivés nés vers la fin du 19e siècle et plus tard», l'autre «le ton de conversation naturelle de la classe cultivée» ; Fouché précise qu'il s'est inspiré des «ouvrages déjà parus» et qu'il a «enquêté lui-même pendant de nombreuses années dans les milieux cultivés de la capitale», et qu'il a «dégagé une moyenne de tous ces documents» (30). Mais Fouché ne pratiquait pas l'analyse quantitative d'un corpus fermé, et loin d'une moyenne «objective», il construit une moyenne qualitative c'est-à-dire normative, prescriptive : il s'agit d'indiquer le bon usage. Delattre ne dit rien de ses sources mais elles semblent du même ordre (ainsi ses exemples d'homophonie sont tirés directement du Manuel *de Nyrop, paru en 1902), même si l'intention descriptive est plus nette et la variation à l'intérieur du «ton» désigné, plus précisément observée. L'un et l'autre se réfèrent à l'état de la langue dans la première moitié du siècle (31) »* (Encrevé 2003 :46).

Toutefois Encrevé précise qu'il ne fait pas d'enquête sociolinguistique à la manière de Labov :

> *« (…) mes observations sont à des fins strictement linguistiques ; (…) je cherche à connaitre des faits linguistiques ; ce qui n'empêche que, même pour ce premier temps, il me parait tout à fait nécessaire d'opérer avec des locuteurs qui soient sociologiquement spécifiés, et de mettre en œuvre d'une façon permanente un contrôle sociologique explicite de la situation d'enquête ». (Encrevé 1978, inédit[111])*

[110] ENCREVÉ, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*.
[111] Notes pour une communication orale au colloque sociolinguistique de Rouen, 29 novembre 1978.

Cependant l'enquête reste primordiale pour ne pas refouler la nature sociale de la langue et c'est en ce sens qu'il conclut son étude sur la LSE chez les hommes politiques de la façon suivante :

« *Notre enquête permet de penser que le phénomène s'est grammaticalisé dans les années 70, au moins chez les professionnels de la parole publique, ce qui a permis à tous les Français de se familiariser inconsciemment avec ces prononciations. Aujourd'hui, la liaison sans enchaînement paraît en pleine croissance parmi l'ensemble des locuteurs fortement scolarisés, et la tendance générale à fixer certaines consonnes finales autrefois «instables» (but, fait, quand) devant consonne indique que des phénomènes divers contribuent ensemble à accroître le nombre des syllabes fermées prononcées à la fin des mots.* »

Retenons pour notre propos, que le fait que les LSE soient réalisés systématiquement sur des liaisons facultatives (LF) plaide en faveur d'une approche différenciée des deux types de liaisons. La LF, et a fortiori les LES, nécessitent une véritable approche sociologique qui doit s'appuyer sur une science des données de l'enquête. Notons également que Pierre Encrevé s'interroge dès le début sur la détection d'un changement linguistique en cours.

Ces quelques éléments extraits des travaux d'Encrevé sont loin de rendre justice à ceux-ci de toute leur richesse mais ils permettront de mieux suivre les discussions présentées dans les chapitres suivants.

3.7.3. L'étude de la liaison dans ESLO : De Jong et la sociophonologie de la liaison orléanaise [stratification sociale] [retour]

Dejong a réalisé une thèse résumée dans un article (De Jong 1994) qui s'appuie sur une reprise des données d'ESLO1.

A partir d'un sous corpus de 45 entretiens comprenant un échantillonnage en 3 groupes d'âge, deux pour le sexe et cinq correspondant aux catégories de l'échelle d'Alix Mullineaux, De Jong a étudié 16 000 contextes de liaisons.

> *« La stratification socio-économique retenue est celle de Mullineaux et Blanc (1982). Elle est basée sur la profession et le niveau scolaire des témoins, et elle compte cinq catégories, dont A est la catégorie la plus prestigieuse, et E la catégorie la moins prestigieuse » (De Jong, 1994: 97)*

Les résultats sont présentés par catégories grammaticales (14)

Catégorie	% lia	N
1. articles	99,9	2347
2. numéraux	98,9	284
3. adj. pronominaux	98,6	507
4. pronoms clitiques	95,9	3898
5. adj. prénominaux	94,3	144
6. adj. indéfinis	93,4	227
7. complémenteurs	91,4	536
8. prépositions	84,6	631
9. spécifieurs adverbiaux	84,0	476
10. pronoms indéfinis	75,5	71
11. *être*	54,9	2555
12. aux. de mode	12,7	474
13. négations	11,3	890
14. aux. *avoir*	4,5	356

Tableau 3. Fréquences de liaison après 14 catégories grammaticales.

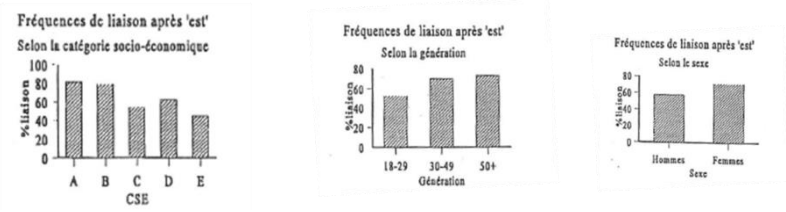
puis à l'intérieur de chaque catégorie, par formes lexicales (ex. : *est* / *était* / *suis* etc.)

Monosyllabique			Polysyllabique		
mot	%Lia	N	mot	%Lia	N
sommes	71.4	28	*étaient*	20.6	34
est	69.0	1692	*étant*	20.0	10
sont	46.0	200	*était*	19.3	212
soient	30.0	10	*serait*	5.6	18
suis	29.2	209	*étais*	5.3	76
soit	10.8	37	*étions*	0.0	9
êtes	0.0	11	*seraient*	0.0	2
es	0.0	1	*serais*	0.0	2
sois	0.0	1	*étiez*	0.0	1
			seras	0.0	1
			seront	0.0	1
total	61.5	2189	total	15.0	366

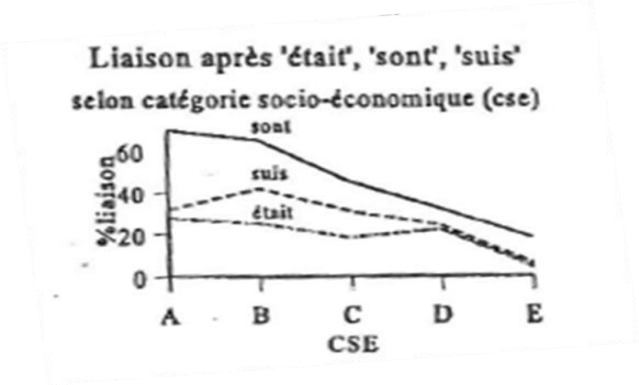
Tableau 4. Fréquences de liaison relatives (%Lia) et absolues (N) après les formes de l'auxiliaire *être*.

et s'appuient sur une corrélation avec les données sociologiques sur les locuteurs (exemple après « est ») :

CSE Age Sexe



Les résultats de De Jong sont significatifs pour une approche sociolinguistique (après « était », « sont » et « suis » selon échelle AM):



Comme le précise Dugua (inédit), on peut résumer l'apport de De Jong a une première étude de la liaison sur un très grand corpus. Il s'attache au taux de liaisons réalisées :

- résultats globaux par catégories grammaticales,
- en fonction de formes lexicales (pour *être* : *est*, *suis*, *était*, etc.),
- en fonction de la consonne de liaison (/z/ vs /t/ par ex)

qu'il croise avec des informations sociales élémentaires : CSP, âge, sexe, ce qui fournit un aperçu général du phénomène. Son objectif est de décrire les aspects phonologiques. C'est ainsi qu'il propose que chaque mot peut avoir deux représentations lexicales supplétives : une représentation sans consonne flottante et une représentation avec consonne flottante. Pour cela il s'appuie de manière innovante sur le rôle des aspects lexicaux (il rapproche la fréquence d'occurrence d'un mot et la fréquence de la liaison) et compare l'acquisition de la liaison à l'acquisition du vocabulaire.

3.7.4 Etude de la liaison et corpus contemporains [retour]

Si pour certains phonologues (Scheer 2004) les corpus n'apportent que très peu à l'étude de la phonologie, différents corpus sont à l'origine d'analyses sur la liaison en français. Nous pouvons même repérer une relation forte entre un domaine et des données empiriques :

Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ? (Baude & Dugua 2015:353-354)

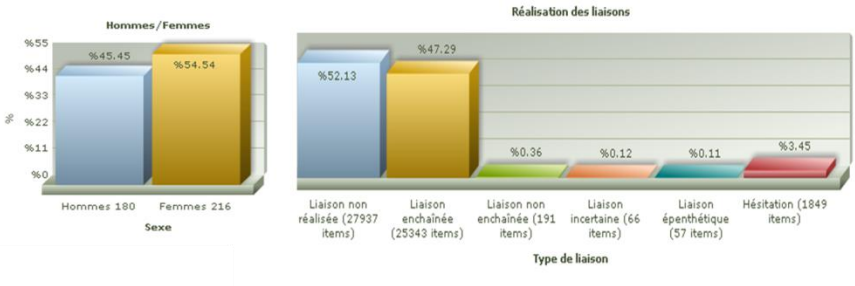
Dès le début des années 1970, Ågren (1973), à partir de 40 heures d'enregistrements radiophoniques impliquant des journalistes, des hommes politiques et des écrivains, fournit un premier travail sur la liaison à partir d'un corpus de parole spontanée. Laks (1980) dans son travail de thèse a constitué un corpus d'adolescents de banlieue parisienne et a ainsi marqué un tournant dans le champ de la sociolinguistique en France. A la même période, Encrevé (1988) a rassemblé les paroles d'hommes politiques afin de rendre compte, notamment, d'un nouvel usage de la liaison : la liaison sans enchainement. Par la suite, d'autres corpus ont été constitués avec pour objectif premier l'étude de la liaison, citons évidemment PFC (Durand et al. 2002 ; Durand et al. 2011), mais également plus récemment le projet ALIPE (Liégeois et al. 2011) ou HPOL (Laks 2007). Dans ces corpus, les contextes de liaisons sont repérés et les réalisations / non réalisations codées. Plus précisément, dans le corpus PFC par exemple, il est possible d'obtenir des résultats généraux facilement, mais aussi d'affiner ces derniers en sélectionnant des critères sociodémographiques (âge, sexe des locuteurs), situationnels (lecture, discussion guidée, discussion informelle), géographiques et de type morphosyntaxique. Les choix de PFC sont clairement énoncés dans leurs publications de présentation du projet : le codage est systématique et ne s'appuie sur aucune classification préalable telle que celles définies dans les ouvrages de références antérieurs. La méthodologie de corpus permet ici de s'affranchir de cadres théoriques afin de réaliser une

première étape purement descriptive de données attestées. Par la suite, l'analyse de ces données met explicitement en question ces classifications (Durand & Lyche 2008).

Le projet ALIPE, quant à lui, fournit un contexte particulier, celui des interactions parents-enfants, et permet d'observer à la fois l'usage de la liaison chez les parents, lorsqu'ils s'adressent à l'enfant ou se parlent entre eux, et chez les enfants (âgés entre 2 et 4 ans).

Les autres corpus de français parlé disponibles – citons par exemple CFPP2000, Valibel, CFPQ, OFROM, Clapi, CoLaJE, ESLO – constituent des bases de données exploitables pour étudier la liaison mais ne fournissent pas d'outil de repérage et d'exportation des contextes de liaison. La mise en perspective des différentes bases de données devrait permettre d'une part, de disposer d'une masse de données importante et d'effectuer des analyses quantitatives et, d'autre part, d'avoir accès à des situations de communication et des locuteurs variés.

Parmi ces projets, PFC affiche les résultats les plus généraux :



Selon Durand et al. (Durand et al.2011 :113)

> *« Considérons d'abord les liaisons attestées. Dans nos données, la liaison est réalisée dans 23 953 cas, soit 47,4 % des sites possibles. Ces sites potentiels incluent donc ceux de liaisons possibles, mais non attestées ou aléatoires dans PFC (par exemple, une liaison entre un SN sujet pluriel et le SV suivant : Les enfants_arrivent). Les liaisons non réalisées sont ainsi la réunion de deux ensembles : les liaisons non attestées (« interdites » dans la tradition prescriptive) et les liaisons variables non réalisées. Une donnée aussi brute se révèle néanmoins intéressante. En effet, si on compare les liaisons réalisées dans les conversations (libres et guidées) et dans le texte, la différence est très nette : en conversation, la liaison est réalisée dans 43,4 % des cas alors qu'en lecture elle l'est dans 59,4 %. Ce résultat corrobore ceux déjà bien connus : la lecture à haute voix induit une montée statistiquement significative du nombre de liaisons réalisées. Nous reviendrons plus loin sur les variations liées au style. »*

L'ensemble de ces analyses reposent sur des études statistiques possibles à l'échelle d'un grand corpus (Durand et al. 2011 :120-121):

« Ces résultats méritent d'être mis en perspective. Si, depuis L. Clédat (1917) au moins, de très nombreux auteurs avaient souligné l'importance de la dimension fréquentielle et l'étroitesse du nombre de constructions impliquées dans la liaison, jamais à notre connaissance une étude quantitative de grande ampleur n'avait pu établir ces faits de façon incontestable. PFC le permet pour la première fois en analysant 372 locuteurs différents localisés en 35 points de l'espace francophone mondial. Ces locuteurs ont réalisé 16 805 liaisons dans les conversations. Dans ce corpus de grande ampleur donc, 21 constructions différentes seulement constituent les sites de 79,4 % des réalisations, et la fréquence cumulative du 227e site atteint asymptotiquement 100 %. Ainsi, la maîtrise d'un très petit nombre de constructions extrêmement fréquentes et d'un lexique très limité suffit à rendre compte de l'usage réel de la liaison en français contemporain. On aura remarqué que ce résultat, remarquable par sa concision, suit et illustre les conclusions de G. Zipf (1935, 1949) : un très petit nombre d'occurrences très fréquentes assume l'écrasante majorité des cas possibles[31]. Dans nos tableaux en effet, comme prévu par cette loi, le produit du rang par la fréquence cumulative tend à être constant. »

Note 31 : Plus précisément, ce que l'on appelle la loi de Zipf affirme que, étant donné un corpus d'énoncés de langues naturelles, la fréquence de tout mot est inversement proportionnelle à son rang dans une table de fréquence. Ainsi, le mot le plus fréquent sera attesté deux fois plus souvent que le deuxième mot dans la table de fréquence et celui-ci quatre fois plus souvent que le troisième mot, et ainsi de suite.

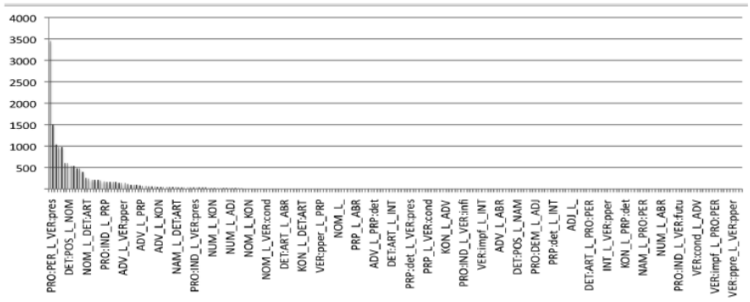


Figure 2 : Fréquence des occurrences pour les 234 contextes grammaticaux dans la réalisation de la liaison orale [a]

a. Pour des raisons d'espace, il n'est pas possible de reporter graphiquement toutes les étiquettes des 234 contextes. Un intervalle de 6 contextes pour une étiquette est donc adopté.

Le corpus PFC, dont une partie des données et résultats sont disponibles sur le site du projet, permet donc un début de comparaison avec d'autres études sur corpus. Ainsi on peut rapprocher les résultats de De Jong et ceux de PFC :

	Nb de contextes	Nb de liaisons réalisées	Nb de liaisons réalisées sans enchainement	Taux de liaisons réalisées avec enchainement	Taux de liaisons réalisées sans enchainement
ont (ont + 'ont)	180	19	0	10,56%	0,00%
sont	181	47	1	25,97%	0,55%
c'est	1477	387	6	26,20%	0,41%
est	628	269	0	42,83%	0,00%
quand	644	442	13	68,63%	2,02%

PFC reste donc exemplaire si l'on considère la quantité de données produites et la multiplicité des analyses effectuées. On peut néanmoins regretter deux aspects : la faiblesse de l'approche sociologique qui a été (nécessairement ?) réduite au profit d'une collecte systématique d'une masse de données et l'absence d'une véritable réflexion sur l'interopérabilité des données, qu'elles soient linguistiques ou sociologiques.

En effet, d'une manière étrange au regard de l'importance de ce type de projet et de l'existence d'autres projets d'envergure dédiés à l'analyse de la liaison, il n'existe pas de « ponts » entre ces différents projets et ces différents corpus. Si certains d'entre eux prennent en compte des objectifs généraux d'interopérabilité (par exemple la TEI pour Alipe, Chat pour Colaje), aucun ne traite cette question directement au niveau de la comparaison d'analyses sur la liaison. Ainsi les codages sont tous différents et la réflexion méthodologique reste fortement contrainte par le périmètre de chaque projet.

Ces lacunes forment les objectifs de l'étude de la liaison dans le corpus des ESLOs : démontrer l'importance de la qualité des données variationnistes, de l'approche réflexive sur leurs conditions de production et la nécessité d'assurer cumulativité et interopérabilité des données.

3.7.5. Une étude exploratoire dans ESLO [retour]

A la différence d'autres corpus, ESLO n'a pas été réalisé pour répondre à un objectif spécifique comme l'étude de la liaison. Afin de savoir si une telle étude était néanmoins pertinente, nous avons (Baude & Dugua 2012, 2013, 2015) mis au point une série de microanalyses afin d'effectuer un « carottage » du corpus :

> Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ? (Baude & Dugua 2015:358)
>
> *Si l'architecture même du corpus (corpus design) des ESLO2 a été conçue dès l'origine pour être comparable avec celle des ESLO1, il convenait de s'appuyer sur des analyses fines du premier corpus afin de mieux maîtriser les effets de transformation de l'objet que peut entraîner une telle méthodologie quantitative.*
>
> *Nous avons donc souhaité, dans une démarche exploratoire et préparatoire, procéder à des micro-analyses à partir de différents axes d'approche du corpus afin de confronter des*

résultats partiels à de grandes tendances. Nous avons interrogé le corpus sur plusieurs extraits, délimités par des critères variés selon une méthodologie de « carottage ». Le terme carottage est utilisé dans un sens métaphorique pour définir une méthodologie empruntée à d'autres disciplines et qui consiste à analyser des échantillons du corpus à travers les différentes strates qui le constituent. Nous considérons en effet chacune de ces micro-analyses fondée sur des échantillons (carottes), comme une possibilité de sonder un grand corpus non pas à différents « endroits » mais plus exactement selon différents angles et par là, différentes perspectives théoriques.

Il s'agit ici de ponctionner quelques carottes à partir de critères issus de postulats variationnistes selon lesquels les « données » ne sont jamais « données » (Encrevé, 1976 : 13) mais relèvent à la fois de contraintes méthodologiques et des outils d'observation et d'analyse mobilisés pour la recherche.

Par commodité, nous avons réparti les « carottes » selon les approches classiques en variation diachronique, diastratique et diaphasique. Nous sommes bien conscients qu'il s'agit là d'une facilité méthodologique qui ne recouvre pas une réalité qu'il revient à l'analyse de prouver : la nature sociale de la langue ne se découpe pas en différents axes ou niveaux de variation.

Variation diachronique

Pour cette première micro-analyse, nous nous sommes appuyés sur un corpus original : celui des sept locuteurs ESLO1 réenregistrés 40 ans plus tard.

Code locuteur	Sexe	Année et lieu de naissance	Age fin d'étude	Age*	Profession*	Echelle AM*
DJ39	H	1932 Amiens	27 ans	37 ans	Médecin ophtalmologiste	A
				73 ans	Retraité (médecin)	A
QB100	F	1945 Orléans	25 ans	24 ans	Élève infirmière	B
				60 ans	Cadre hospitalier	A
CF4	H	1943 Tillay le Peneu	17 ans	26 ans	Ajusteur	C
				64 ans	Retraité (ajusteur)	C
PY94	F	1943 NR	14 ans	26 ans	Gardienne d'immeuble	D
				64 ans	Retraité (contrôleur PTT)	C
RF211	H	1947 Entre Deux Guiers	25 ans	22 ans	Elève professeur	A
				58 ans	Professeur de physique	A
YR399	H	1942 Limoges	17 ans	27 ans	Rectificateur	D
				65 ans	Retraité (ouvrier qualifié)	C
YT387	H	1928 Saumur	18 ans	41 ans	Contre maitre outilleur	C
				77 ans	Retraité (Contre maitre outilleur)	C

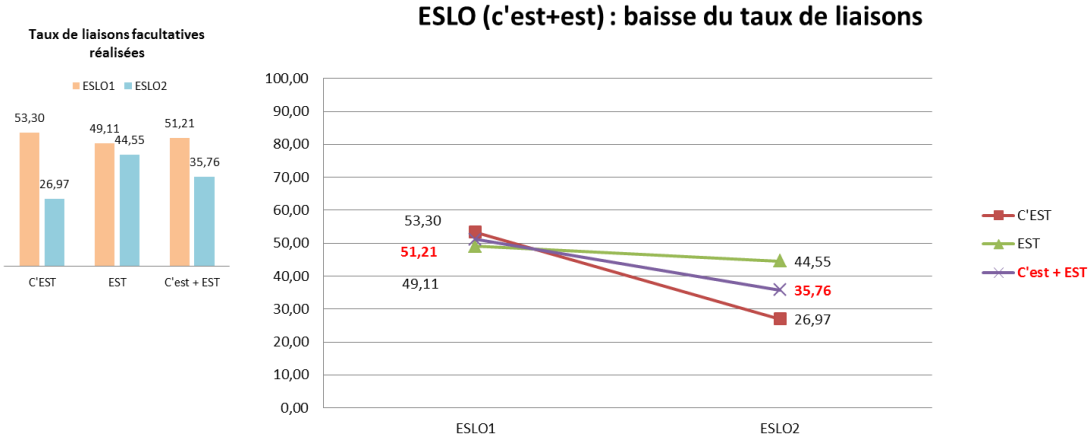
Tableau : Informations concernant les locuteurs du module DIAchronie

* Pour les colonnes Age, Profession, Echelle AM, les deux lignes par locuteur correspondent aux informations au moment d'ESLO1 et au moment d'ESLO2

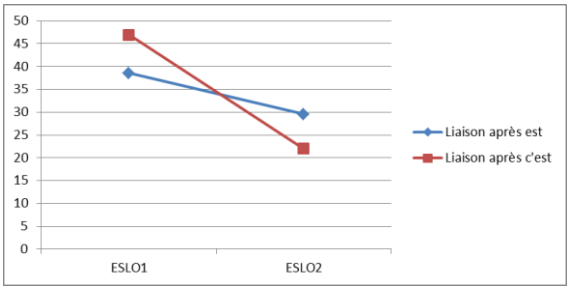
Ce sous corpus est constitué de 17 heures et 11 minutes d'enregistrement (9 heures et 22 minutes d'ESLO1 et 7 heures et 49 minutes d'ESLO2 — 89 625 mots pour ESLO1 et 70 924 pour ESLO2).

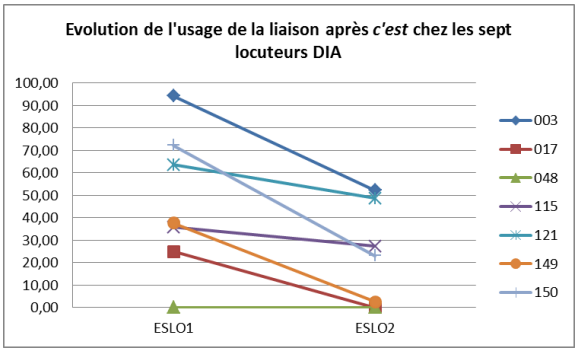
Nous avons restreint notre étude aux usages de la liaison après « c'est » et après « est ». Précisons que la locution « c'est-à-dire » à été traitée à part.

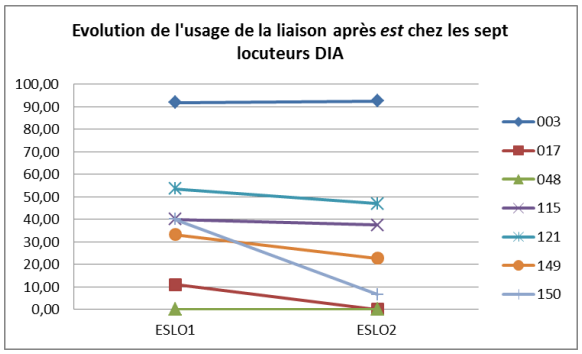
Le premier résultat est de constater une baisse générale du taux de liaisons :



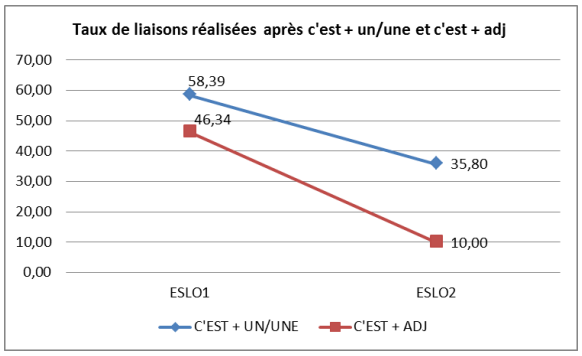
La baisse est plus forte après *c'est* (baisse de 25 points en moyenne) qu'après *est* (baisse de 9 points en moyenne).





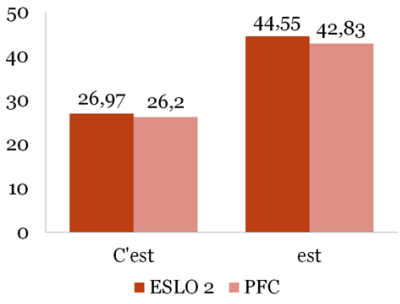


Nous n'analyserons pas ici ces différences. On peut toutefois se demander s'il existe un effet de co-occurrence qui pourrait expliquer la réalisation des liaisons après *c'est* ? En effet, si on distingue *c'est + un/une* et *c'est + adj* on constate que globalement, la liaison est plus souvent réalisée dans le contexte *c'est + un/une* (50.8%) que dans le contexte c'est + adj (34.4%).



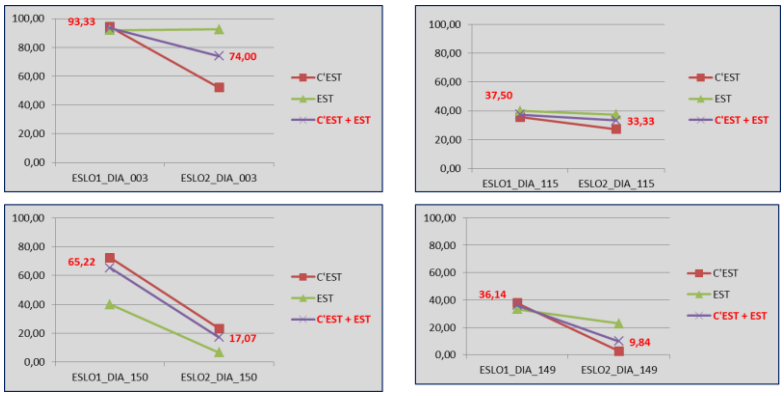
Une façon d'expliquer ce type de résultats est de dire que le fonctionnement de la liaison est plus figé dans des co-occurrences mot1-mot2 fréquentes. Cette piste semble suivie dans des travaux sur l'acquisition : les liaisons obligatoires sont apprises plus tôt dans des co-occurrences mot1-mot2 fréquentes que dans les adjacences rares (Chevrot, Chabanal, Dugua, 2007, p.113-115 et, pour PFC, Côté, 2015[112]).

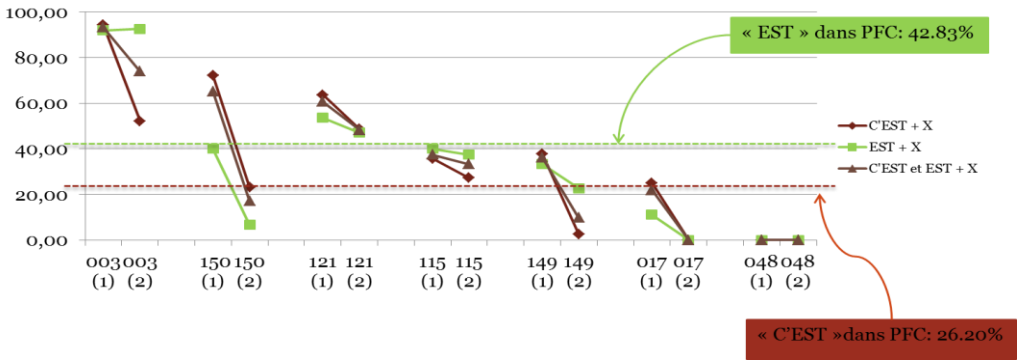
Les résultats généraux sont de fait compatibles avec d'autres études. Une rapide comparaison avec les données PFC montre la cohérence des premiers résultats, et par-delà valide les conclusions tirées de l'analyse de cette première carotte :



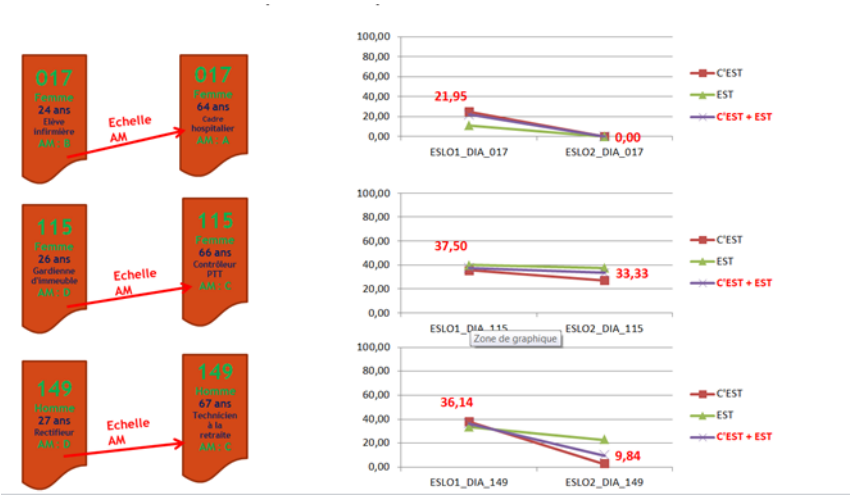
[112] Communication lors du sémianaire « Liaisons » Modyco, Paris , 2015.

L'observation fine des productions des sept locuteurs permet de nuancer ce premier résultat. Les figures ci-dessous montrent des taux de réalisations dans ESLO1 variant entre 0% et 93% et dans ESLO2 entre 0% et 74%, respectivement pour les locuteurs 048 et 003. Entre ces deux extrêmes, on constate une répartition relativement homogène pour les cinq autres locuteurs. Outre des variations dans les taux de réalisation, on peut observer des courbes d'évolution variées. Une pente forte pour le locuteur 150, des pentes comparables (d'environ 20 points) pour les 003, 017, 149, une stabilité pour les 115, 048.





Afin d'approfondir l'analyse de ces variations, nous pouvons utiliser l'appareillage sociologique d'ESLO, la finesse de description des locuteurs, la classification AM et les métadonnées situant l'ensemble des données. Nous avons ainsi regardé plus précisément le parcours de trois locuteurs qui ont effectué une progression sociale selon l'échelle AM :



Usages de la liaison dans le corpus des ESLOs : vers de nouveaux (z)ouvrages de référence ? (Baude & Dugua 2015:363-64)

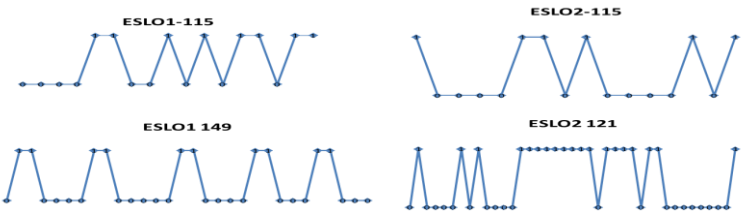
Une combinaison des critères diachroniques et diastratiques permet de repérer des différences de comportement significatifs. En effet si la baisse du taux de liaison est bien présente chez les trois locuteurs conformément aux prévisions des statistiques sur l'ensemble des corpus, nous constatons (voir figure) que le locuteur YR399 effectue une baisse de 27 points, le locuteur QB100 de 22 points et le locuteur PY94 de 4 points seulement. Par ailleurs, ces différences ne compensent pas des variations importantes, puisque les trois locuteurs conservent un taux de liaison très différent (respectivement 9,84%, 0% et 33,33%). Enfin, nos présentes données vont à l'encontre des tendances générales sur l'usage de la liaison en fonction des données sociales (Ashby 1981, De Jong 1991) : c'est ici la locutrice qui se situe dans l'échelle la plus élevée qui a le taux de liaison le plus bas. Il faudrait prendre en compte d'autres critères pour cerner cette variation et s'interroger sur les conditions d'un éventuel changement linguistique. Ainsi, Lyche et Otsby (2009), dans leur article consacré au français de la haute bourgeoisie parisienne, relèvent que « (…) toutes choses égales par ailleurs, certains locuteurs sont plus susceptibles que d'autres de produire des liaisons et les locuteurs les plus âgés exhibent un taux de liaisons plus conséquent *».*

Cette analyse est bien évidement à prendre avec la plus grande prudence puisqu'elle se fonde sur une analyse quantitativement négligeable. Toutefois elle nous semble suffisamment significative pour nécessiter une étude approfondie en croisant les critères diachroniques, diastratiques et diaphasiques.

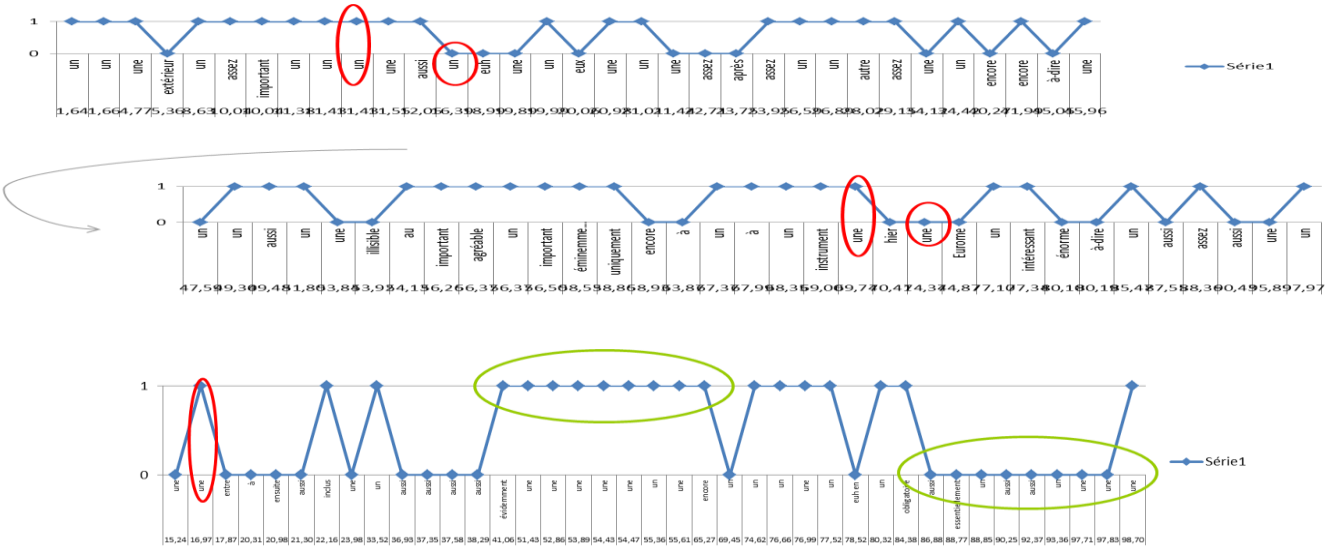
Les variations d'un même locuteur selon les situations de communication ont été décrites depuis Labov 1973. Il nous a donc semblé pertinent d'analyser la situation d'entretien au

regard de l'approche réflexive menée dans le cadre de l'élaboration du protocole ESLO2 et de la confronter à une comparaison avec d'autre situations.

Dans un premier temps nous avons réalisé une simple « carotte » afin d'étudier les variations d'un même locuteur au sein d'un entretien.
Le graphique ci-dessous montre l'extrême variabilité de ce critère :



Les graphiques suivants détaillent l'entretien 121 :



Un premier constat permet de repérer des variations qui, si elles sont loin d'être aléatoires, ne sont pas pour autant facile à analyser à l'aide des outils classiques appliqués sans tenir compte du contexte. Il est par exemple convenu que le début d'un entretien représente un moment où le locuteur se surveille particulièrement et nombre d'études écartent de l'étude cette partie de l'enregistrement. Cette méthode est loin de suffire pour écarter des biais importants.
Derrière ce premier constat nous ne pouvons que saisir un encouragement à une étude qualitative qui prenne en compte l'ensemble des données disponibles afin d'intégrer toutes les sources de variation. Nous poursuivons donc notre étude avec un focus sur une analyse qui croise variation diachronique, diastratique et diaphasique. Notons que celle-ci n'est possible que compte tenu des choix de constitution et de structuration des données, situées et contextualisées, dans ESLO.

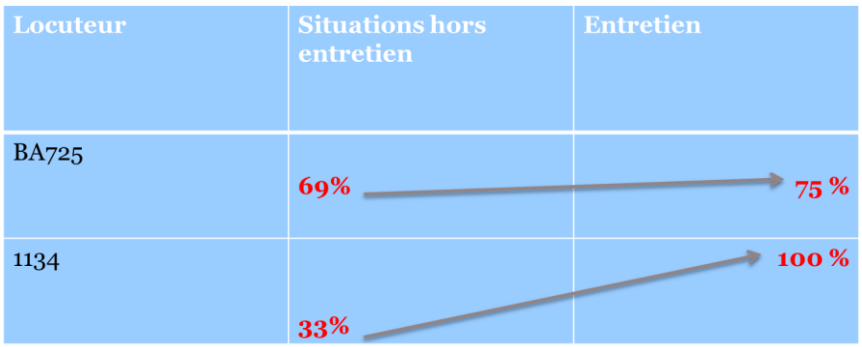
Afin de réaliser cette nouvelle micro-analyse, nous avons constitué un sous corpus de quatre locuteurs (deux de chaque sexe, appartenant à des catégories différentes de l'échelle AM : A, B, D) enregistrés dans des situations variées. Le total représente vingt-quatre enregistrements, de durée variable (de 1 à 89 minutes), pour un total de 11h08 minutes (environ 100 000 mots).
Une première analyse conforte les analyses antérieures de De Jong :

	De Jong 1994	Baude, Dugua 2011
est+X	69 %	65 %

après *(c')est*, nous constatons un taux de réalisation de 65%, comparable aux 69% comptabilisé par De Jong (1994). Toutefois une analyse qui combine les différents critères de variations tout en intégrant des comportements décelables à un niveau plus fin est riche d'enseignements et bouleverse quelque peu cette impression d'homogénéité.
On restreint le corpus aux deux locuteurs hommes qui selon les critères classiques de la sociolinguistiques seraient classés dans la même catégorie :

- Gilbert (BA725), 58 ans, vendeur, certificat d'études, fin d'études à 14 ans, AM = D
- Georges (1134), 58 ans, vendeur, certificat d'études, fin d'études à 13 ans, AM= D

Un comptage des liaisons après *(c)est* dans deux situations différentes démontre des comportements linguistiques différents :



Si on les regroupe, les taux de Gilbert et Georges sont comparables, mais si on les distingue, on constate que Gilbert à un taux relativement stable avec une légère augmentation lors des entretiens alors que Georges fait peu de liaisons facultatives hors entretien et beaucoup (toutes) lors de l'entretien.
Comment expliquer ces variations ?
Si on prend en compte la méthodologie de classification d'Alix Mullineaux dans la version qu'elle souhaitait développer – c'est-à-dire en prenant en compte la capital culturel et la trajectoire sociale, comme cela est attendu dans ESLO2 – et les informations qui permettent

de contextualiser les conditions de production des données (rôle de l'interviewer, objectif annoncé, mode d'approche etc.), nous pouvons faire parler les données. En effet, pour résumer les informations complémentaires, nous pouvons décrire la trajectoire de ces deux vendeurs dont l'un seulement changera de « catégorie AM » :

- Gilbert souhaite devenir « *gérant de plusieurs boucheries* », il aime « *la lecture et la musique* », compte « *visiter des musées quand [il sera] à la retraite* ». Il deviendra gérant de plusieurs boucheries dont la plus prestigieuse de la ville mais aussi des toutes nouvelles boucheries de supermarché. Nous pouvons donc décrire une trajectoire qui va d'un indice D (C ?) à un indice B de l'échelle AM (version complète).
- Georges *« rêvait d'être boulanger, ne prend pas de vacances sauf la pêche* », a été « *une fois au cinéma en 17 ans* » et ne connait pas « *le dictionnaire utilisé par [son] enfant »*. Pas d'évolution professionnelle et pas de trajectoire ascendante.

Précisons que Gilbert est interviewé par un homme d'une cinquantaine d'années après une longue phase de présentation de l'enquête, entretien réalisé au début de l'enquête, alors que Georges est interviewé par une jeune femme, entretien réalisé lors de la période intensive de l'enquête avec une explications du projet différente. Il s'agit de deux modes d'approche distincts dont la description est très difficile à reconstruire si elle n'est pas documentée (i.e. si elle ne fait pas partie intégrante des données).

Ces informations, croisées avec l'analyse des taux de liaison, permettent de déceler des stratégies linguistiques bien plus prononcées chez Georges qui fait un effort important pour passer de 33% à 100% de liaisons réalisées en situation formelle (entretiens). L'habitus linguistique de Gilbert est différent et son taux de liaison est moins sensible à la variation diaphasique mais aussi plus élevé d'une manière générale**. On est ici au cœur de la variation linguistique, que seules l'enquête sociolinguistique et l'exploitation maitrisée du corpus permettent d'atteindre.**

Nous avons ici, concentrées, toutes les données pour comprendre un changement linguistique. La différence entre le taux de LF de Gilbert et Georges en entretien annonce la baisse constatée sur l'ensemble des corpus mais la stabilité affichée par Gilbert dans toutes les situations par rapport à Georges qui use de stratégies linguistiques démontre que les liaisons restent un marqueur social important. Si la baisse est effective, elle ne peut atteindre un niveau trop bas et elle rend surtout compte d'un nivèlement superficiel du côté formel de certaines situations. Pour l'imager d'une métaphore simpliste, ce n'est pas parce que la cravate perd son critère d'accessoire obligatoire en situation formelle que le dégrafé de la chemise ne remplace pas avec plus de discrétion un lieu de lutte sociale identifiable par les agents.

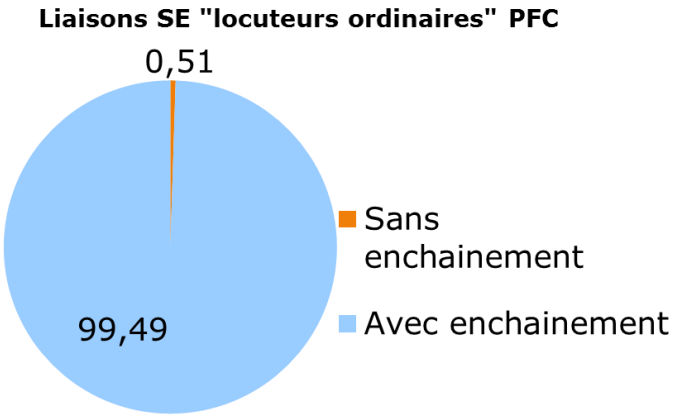
Avec cette micro-analyse, nous confirmons l'objectif d'ESLO en tant que réalisation d'observables dont les conditions de production sont incorporés à la méthodologie et à l'analyse dans un même mouvement. S'il fallait résumer le travail d'analyse présenté dans cette HDR, cet extrait en serait l'élément le plus significatif.

3.7.6. ESLO et la Liaison sans enchainement [retour]

La découverte de la liaison sans enchainement (LSE) et son analyse par Pierre Encrevé représente un moment clé pour la sociolinguistique et la linguistique en général. Nous ne reviendrons pas ici sur ces travaux et leur importance mais nous l'appréhenderons, c'est le sens de ce travail, par le côté des « observables ».
En effet, si la LSE a donné et donne lieu à des travaux majeurs en phonologie, elle a été peu travaillée par les sociolinguistes. Même le projet PFC qui affiche son ancrage sociolinguistique n'a pas développé d'étude qui permettrait de confronter les travaux d'Encrevé sur la LSE chez les hommes politiques aux pratiques de toute une communauté linguistique.

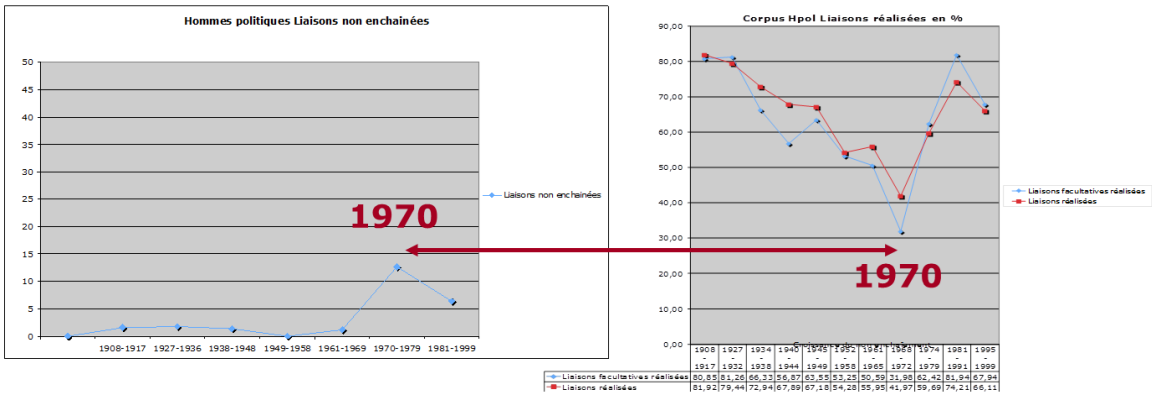
Ainsi, fort de leur très grande quantité de données, les auteurs de PFC font remarquer l'absence de LSE dans leur corpus :

		Liaisons Z	Liaisons T	Liaisons N	Liaisons pkh	liaisons total
Non enchaînées	N Mots	17	13	5	3	38
	N Occurrences	24	43	6	3	76
Liaisons total	N Mots	144	72	43	11	270
	N Occurrences	6682	2730	5272	96	14780



Ils en déduisent que l'hypothèse d'un changement linguistique en cours, repéré chez les hommes politiques, locuteurs légitimes, serait erronée car la LSE, après un siècle de présence dans leurs discours, n'est pas reprise dans les énoncés de conversations ordinaires.
Il est d'ailleurs intéressant de noter que Laks, co-auteur de PFC, a constitué un autre corpus (HPOL) sur les hommes politiques pour étudier l'évolution de la LSE. Il s'agit d'une entreprise qui représente un lourd travail dont il résulte une analyse particulièrement intéressante sur la variation diachronique de la LSE.
Sur presque un siècle de données, il démontre que le taux de LSE, qu'on savait très variable d'un homme politique à l'autre et même chez le même locuteur, reste relativement élevé en

moyenne (11,8%) mais il constate surtout une grande variabilité et un pic à la fin des années 70 qui apparait conjointement à une baisse significative de la LF :



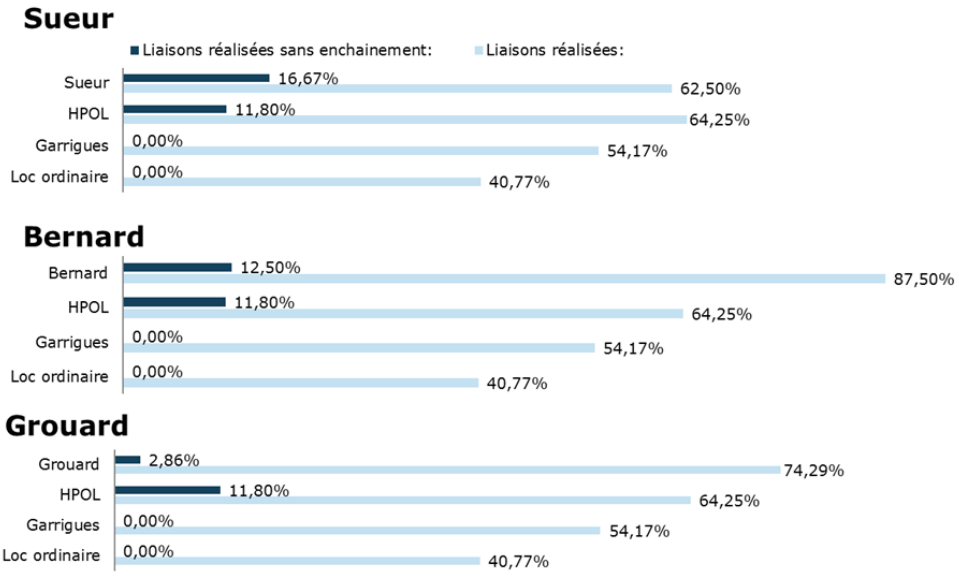
> « *Tout se passe comme si l'apparition des LF non enchaînées au dessus de 10% était liée à une réalisation des facultatives de l'ordre de 30%. Lorsque les facultatives retrouvent un étiage à 60%, les LF non enchaînées baissent en dessous de 10%.* » (Laks 2011)

D'une manière étonnante, cette analyse pertinente reste pour Laks dépendante du corpus HPOL et ne rentre pas dans les éléments d'analyse du corpus PFC. Ces deux corpus semblent indépendants au point qu'on peut se demander s'il ne traite pas de deux communautés linguistiques distinctes. A l'inverse nous pensons qu'un corpus suffisamment représentatif d'une communauté linguistique, a fortiori d'une communauté d'auditeurs, doit permettre une étude qui confronte la parole des discours politique et celle des conversations ordinaires dans un objet regroupant ces différentes pratiques. C'est une perspective qui est en cours de réalisation dans ESLO2.

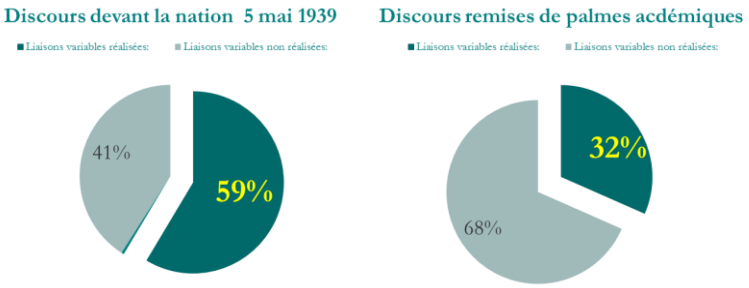
Dans un premier temps, nous avons entrepris de constituer un module discours de représentants politiques dans ESLO2. Les premiers éléments concernent des enregistrements de discours de trois maires d'Orléans[113] (Jean Pierre Sueur, Jean-Louis Bernard et Serge Grouard). Nous avons comptabilisé le taux de LF et de LSE pour chacun de ces maires et nous les avons confrontés aux taux moyen d'ESLO2 et à ceux d'un locuteur familier de la parole publique mais qui n'est pas un élu (Jean Guarrigues, professeur d'université, politologue dans les médias).

L'analyse confirme la présence variable de LSE uniquement chez les hommes politiques :

[113] http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO2_DISC_1237 ou ark:/87895/1.17-475420
http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO2_DISC_1238 ou ark:/87895/1.17-475421
http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-ESLO2_DISC_1239 ou ark:/87895/1.17-475422



Un autre élément du module concerne des enregistrements du plus célèbre homme politique d'Orléans : Jean Zay. Il y a trop peu d'enregistrements disponibles pour procéder à un comptage de la LSE chez Jean Zay, mais deux extraits sont suffisants pour une étude sur la LF et pour relever son extrême variation :



Ces différents éléments nous semblent plaider pour la nécessité de disposer d'un continuum de pratiques linguistiques : il convient de disposer de différents enregistrements des représentants politiques, et pas seulement dans des discours officiels. De même des enregistrements de locuteurs « ordinaires » qui deviennent rompus à la parole publique devraient compléter le portrait sonore d'une communauté. En ce sens, ESLO semble, à partir d'un réel portrait sonore d'une ville, offrir l'opportunité de saisir les pratiques linguistiques diversifiées à la fois des représentants politiques et d'autres agents avec toute la précision requise.

De plus, ESLO permet comparer avec les enregistrements d'autres locuteurs en situation de parole publique et de parole privée avec un degré de précision que seul un projet à la taille d'une ville moyenne permet. C'est, nous semble-t-il, tout l'enjeu d'une approche variationniste sur corpus.

En effet, dès 1983, dans ses analyses sur les conditions de réalisation des liaisons facultatives, Encrevé avait relevé la difficulté de saisir les causes de cette variation :

> « *Que l'écheveau des déterminants de la variation concernant le taux de réalisation des liaisons facultative apparaisse très emmêlé ne signifie pas qu'il faille renoncer à la relier aux caractéristique sociales des locuteurs, alors que tous les faits évoqués semblent pouvoir, au contraire, trouver là, un par un, une explication. Mais pour en rendre compte globalement, il serait nécessaire d'élaborer un modèle très complexe articulant systématiquement toute une série d'informations sociales. S'agissant d'un point explicitement «manipulé» par le système d'enseignement comme l'est la liaison, il faudrait certainement, outre l'origine sociale et le capital scolaire, prendre en compte la notion de génération, entendue principalement au sens de mode de génération, c'est-à-dire non comme classe d'âge biologique mais historique. Une génération réunit des agents ayant participé (différemment) à une même histoire sociale : s'agissant de l'acquisition de la langue, une même génération de locuteurs a été exposée à un même état du système d'enseignement (défini par le mode de recrutement et de formation des enseignants, par le mode d'accès des enseignés à tel niveau d'enseignement, par la durée de la scolarité obligatoire, etc.)* ». (Encrevé 83:60)

L'approche réflexive du projet ESLO met en exergue cette volonté de construire un observable suffisamment complexe et d'en interroger les conditions de production afin de permettre des analyses à partir de « l'écheveau des déterminants de la variation ». Le module « responsables politiques » en cours de collecte consiste donc à enregistrer des personnalités politiques au sein d'un réseau de productions linguistiques appréhendable. Ceci est rendu possible par la maniabilité de l'objet ville. Cela serait difficilement possible dans un cadre national, mais à l'échelle d'une agglomération, il est possible d'enregistrer une personnalité dans de nombreuses situations (discours, interviews, discussion, réunions et même conversations ordinaires). Enfin il est possible de situer cette personnalité dans les différents réseaux représentés dans le corpus.

Seul un tel corpus nous semble apporter les garanties d'une suffisante représentativité de pratiques aussi difficiles à cerner que l'usage de la LSE. C'est donc un vaste chantier pour ESLO qui, nous l'espérons, permettra d'éclairer les travaux sur l'hypothèse d'un changement linguistique en cours ou tout au moins d'une présence accrue de la LSE dans les pratiques d'une communauté linguistique.

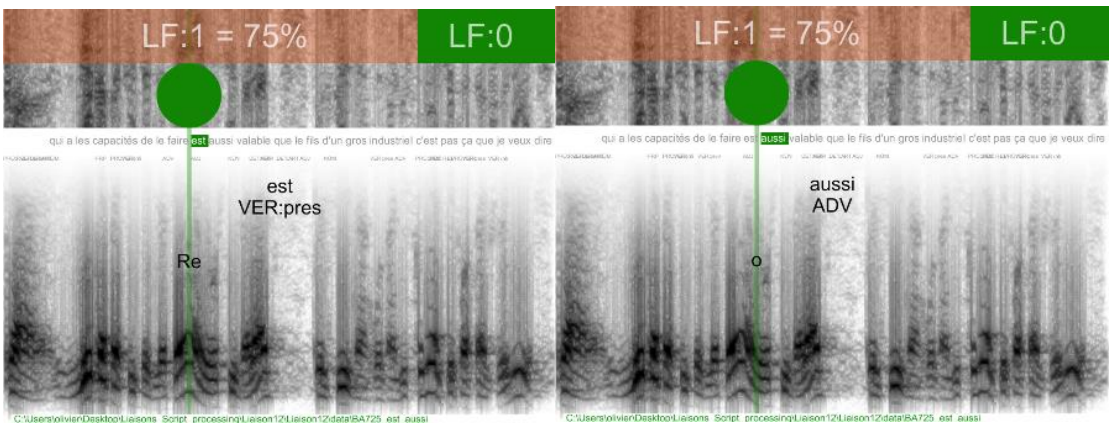
Nous pourrons alors confirmer les réponses aux questions d'Encrevé 1983 et ses propres assertions :

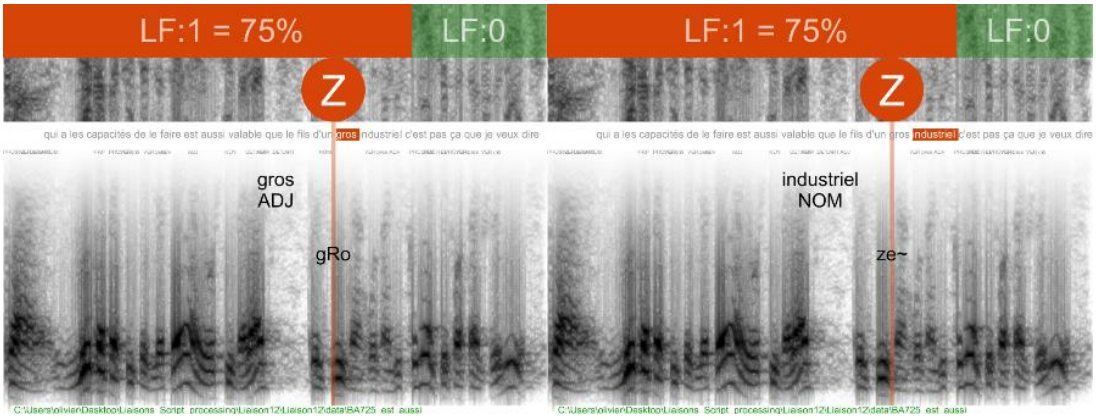
> « *Cette phonologie rend compte de la compétence utilisée en locution par les dirigeants politiques actuels, et, au-delà, par tous les locuteurs réalisant*

des liaisons facultatives, même s'ils ne font entendre qu'exceptionnellement des liaisons sans enchainement. Vaut-elle également pour les locuteurs, qui tels les adolescents témoins de Villejuif ne réalisent qu'exceptionnellement des liaisons facultatives, et jamais de liaisons sans enchainement ? Le traitement linguistique ici proposé vaut donc pour l'ensemble des locuteurs — la différence entre la grammaire des dirigeants politiques et celle des adolescents de Villejuif quant à la liaison portant seulement sur le rattachement de CL à l'attaque flottante, qui est variable chez les premiers, invariable chez les seconds, du moins en production.(Encrevé 1983 :59).

3.7.7. Prolégomènes à une data visualisation de la liaison dans les corpus complexes [conclusion] [retour]

Dans un autre chapitre, nous plaidons pour un développement de la data visualisation afin d'accompagner le chercheur dans l'exploration outillée de données massives et complexes. Les premières analyses d'un phénomène aussi précis que celui de la liaison confirment la nécessité de prendre en compte de nombreux facteurs dans le traitement des données.
Une première expérience très rudimentaire a été testée à partir d'un projet de valorisation du corpus ESLO. Dans le cadre de ce projet, un travail de visualisation artistique à partir du logiciel processing avait ouvert des perspectives intéressantes.
L'objectif de cette ébauche est d'élaborer un outil qui permette, lors de l'écoute d'un enregistrement, de repérer automatiquement les liaisons et d'attirer l'attention sur une réalisation particulière, notamment par le croisement de statistiques et d'analyses du signal acoustique.
La chaine de traitement est la suivante : l'enregistrement est transcrit et le script easyalign est utilisé pour une transcription automatique en sampa, une transcription en API, un découpage et un alignement au mot, à la syllabe et au phonème. Une annotation supplémentaire est ajoutée à partir d'un codage de réalisation de la liaison.
Le textgrid praat est récupéré et converti au format XML. Ce dernier peut alors être lu par un sketch processing développé par Gérard Paresys et Guy Kayser à partir d'un projet développé dans un cadre de R&D.
Les mêmes auteurs ont adapté le script lors d'une collaboration sur le projet liaison. Le sketch permet la lecture du fichier son avec une synchronisation sur la transcription et un traitement des annotations. Lorsqu'un contexte de liaison est traité, celui-ci est signalé en rouge s'il est conforme avec le taux de réalisation habituel du locuteur et en vert s'il s'agit d'un cas marqué. Dans l'exemple suivant, le locuteur réalise les LF dans ces contextes à hauteur de 75%. Dans le premier cas, la LF n'est pas réalisée et le rond vert attire l'attention. Dans le second cas la LF est réalisée et la CL est indiquée comme information supplémentaire.





Cette première ébauche est très rudimentaire mais elle a le mérite de lancer des pistes d'exploitation d'une visualisation dynamique à partir de facteurs multi-variés. Elle n'a d'autre prétention que d'apporter une contribution à une exploration de corpus oraux à partir de l'objet principal : l'enregistrement de la parole en lien avec les différentes données, métadonnées et annotations disponibles dans des corpus complexes.

3.8 Perspectives pour les ESLOs

3.8.1. ESLO 3.0 (2014-2017) [retour]

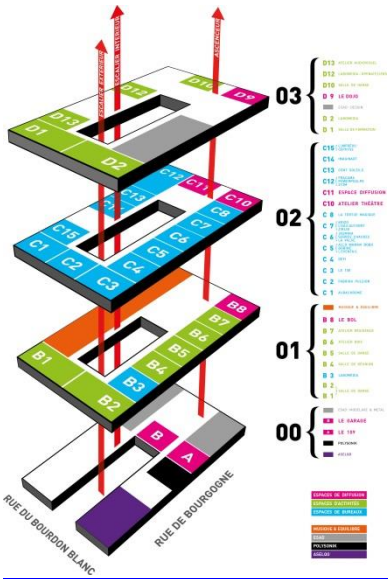
ESLO 3.0 est un projet exploratoire qui anticipe les nouvelles questions de collecte et de traitement des données à l'ère de l'Internet 3.0, c'est à dire d'un réseau qui se tourne vers son aspect contributif et vers la sémantisation des données à partir de leur mise en relation dans des entrepôts structurés.

Dans l'élargissement des configurations d'enquête destinées à étendre les formats d'interaction discursive – en convergence avec l'approche développée à l'Institut für Deutsche Sprache (IDS – Mannheim) dans la perspective de constitution d'un Laboratoire Européen Associé –, le projet ESLO 3.0 souhaite constituer la première expérimentation de collecte et d'exploitation de données sonores collaboratives. Le développement a été conçu suivant trois phases afin d'assurer au programme une maitrise entière du processus dans la collecte et l'exploitation, avant la phase d'expérimentation numérique.

Phase 1. Collecte de données : Portrait sonore du « Collectif 108 »
Le collectif 108 – Maison Bourgogne regroupe en un même lieu des associations orléanaises représentatives des pratiques culturelles et des activités de loisir d'une ville de cette dimension dans les années 2010 (centre de loisirs, spectacles vivants, école de musique, animation sociale, création multimédia etc.).

Présentation du « 108 » :

Le collectif 108 – Maison Bourgogne est une association 1901 regroupant une trentaine d'associations professionnelles dans le secteur culturel : http://le108.org/blog/



Le choix du lieu et des agents répond à un corps d'hypothèses sociologiques sur l'émergence des nouveaux acteurs du changement linguistique et sur les réseaux de diffusion des innovations lexicales et discursives.

Pour commencer, une série d'enquêtes s'est assigné pour objectif de dresser un portrait sonore de la communauté gravitant dans l'immeuble du 108 en se conformant à la méthodologie éprouvée du programme développée par le projet des ESLO. En complément des enregistrements déjà réalisés dans l'agglomération d'Orléans, plusieurs dizaines d'heures, représentatives de la diversité des pratiques linguistiques dans un contexte sociologiquement identifié, seront recueillies sous forme d'entretiens, de réunions, de conversations, de spectacles et de productions, de conférences, complétées par des scènes de la vie quotidienne et des captures d'ambiance. Ce travail s'appuie sur une démarche ethnométhodologique à travers la participation des chercheurs à la vie du lieu et à leur implication dans la communauté des acteurs.

Les enregistrements complèteront le corpus des ESLO en suivant la chaine de traitement mise en place par le laboratoire. Au-delà de l'incrémentation de la base de données, l'objectif poursuivi est de disposer d'un échantillon linguistique de dimension conséquente d'agents impliqués dans la redéfinition des conduites culturelles, dans leur façon de s'y impliquer et dans la relation discursive et réflexive qu'ils entretiennent à leurs pratiques et à leur environnement. La démarche est innovante en ce qu'elle repose sur la recherche de la conjonction qui peut s'établir entre l'exploitation des données et la phase de collecte. A ce titre, les contraintes et les potentialités du traitement sont prises en compte dans les modalités de l'acquisition des ressources et réciproquement. Une thèse est en cours préparée par Athéna Dupont qui se consacre en particulier à l'association Labomédia.

Liste des enregistrements prévus :

- Labomédia (Laboratoire multimédia numérique) :
  - 3 réunions de travail,
  - 12 entretiens,
  - 8 séances atelier accueillant du public
- Compagnie théâtre Zirlib
  - 6 entretiens
  - 3 réunions de travail
  - 1 captation de répétition
- Accueil
  - 6 demi-journées d'interactions hôtesse d'accueil – visiteurs
- Tortue magique (Compagnie spectacles de marionnettes)
  - 3 entretiens
  - 4 captations de spectacles
  - 12 interactions spontanées entre enfants
- Enregistrements pris sur le vif
  - 12 discussions lors d'une manifestation publique

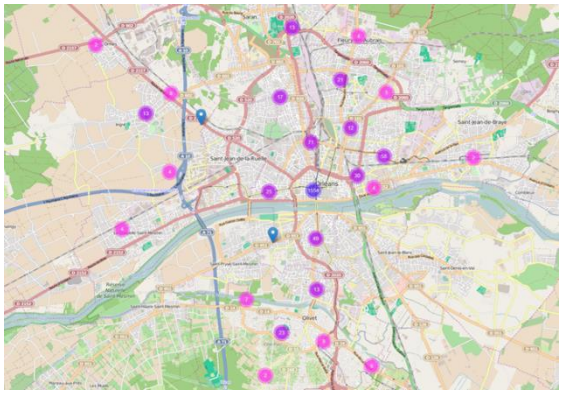
    - 12 discussions « dans les couloirs »
    - 12 discussions lors de la visite d'une exposition
    - 6 captations de présentation d'œuvres lors de manifestation publique ou d'une exposition
- L'ASELQO, organise et anime des activités d'intérêt social dans le domaine culturel, socioculturel et de loisirs, elle accompagne aussi les jeunes porteurs de projet culturel et/ou citoyen. Leurs bureaux, situés à l'accueil de la Maison Bourgogne, se transforment au gré des saisons en lieu d'exposition, guinguette et fêtes à thème.
    - 50 enregistrements divers :
        - Réunions
        - Discussions
        - Conférences
        - Conversations pendant des ateliers manuels
        - Activités ludiques

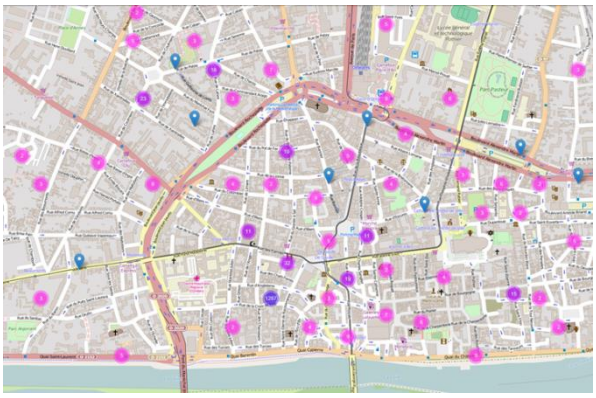
Phase 2. Exploitation partagée des données et visualisation

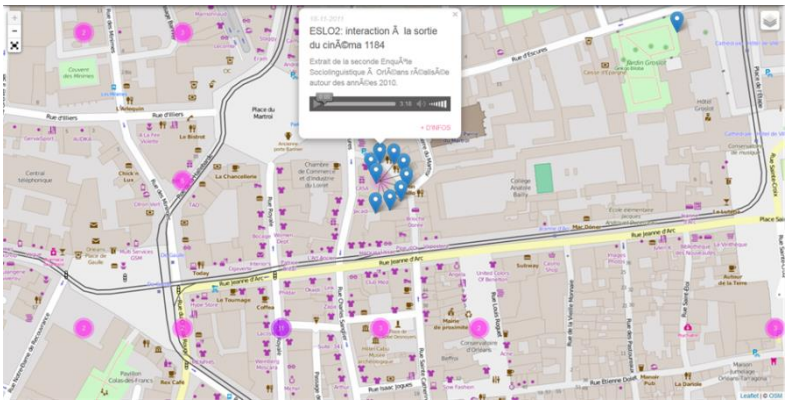
A partir des données, l'investigation porte sur la conception d'un outil de visualisation adapté au corpus ESLO afin de restituer et de figurer les contenus sonores. Ce volet du projet, qui s'inscrit dans le déploiement de nouvelles collaborations entre la recherche et la culture, est à concevoir sous forme d'un partenariat entre le LLL, l'atelier numérique de la MSH et l'association de création numérique Labomédia qui est l'un des acteurs impliqués dans le collectif « 108 ». L'interface de visualisation est élaborée en fonction des données et des métadonnées du corpus ESLO, en exploitant notamment les ressources offertes par la géolocalisation et l'édition graphique de contenus.

D'un point de vue scientifique, il s'agit d'expérimenter les capacités de restitution et de compréhension des dynamiques linguistiques qui s'exercent dans une ville à partir d'un croisement des métadonnées sociologiques et spatiales. Culturellement, l'objectif poursuivi est une présentation et une appropriation des données scientifiques à des fins artistiques et collaboratives et la restitution d'une navigation sonore à partir d'une carte interactive qui permette un accès à des utilisateurs non experts.

La réalisation de la carte interactive se fait à partir des métadonnées de géolocalisation disponibles dans la base de données ESLO (développement en cours) :







Phase 3 : Collecte 3.0

La troisième phase du projet est à concevoir comme un effet en retour sur le processus à partir des résultats obtenus. Alors que dans la phase 1 l'enquête était principalement liée aux compétences des chercheurs, l'obtention d'une interface de visualisation conviviale autorise un maniement interactif, de la part des témoins comme des utilisateurs, et la mise en place subséquente de fonctionnalités de collecte et de descriptions collaboratives. Cette interface sera complétée par la réalisation d'un objet virtuel dédié aux mêmes fonctionnalités.

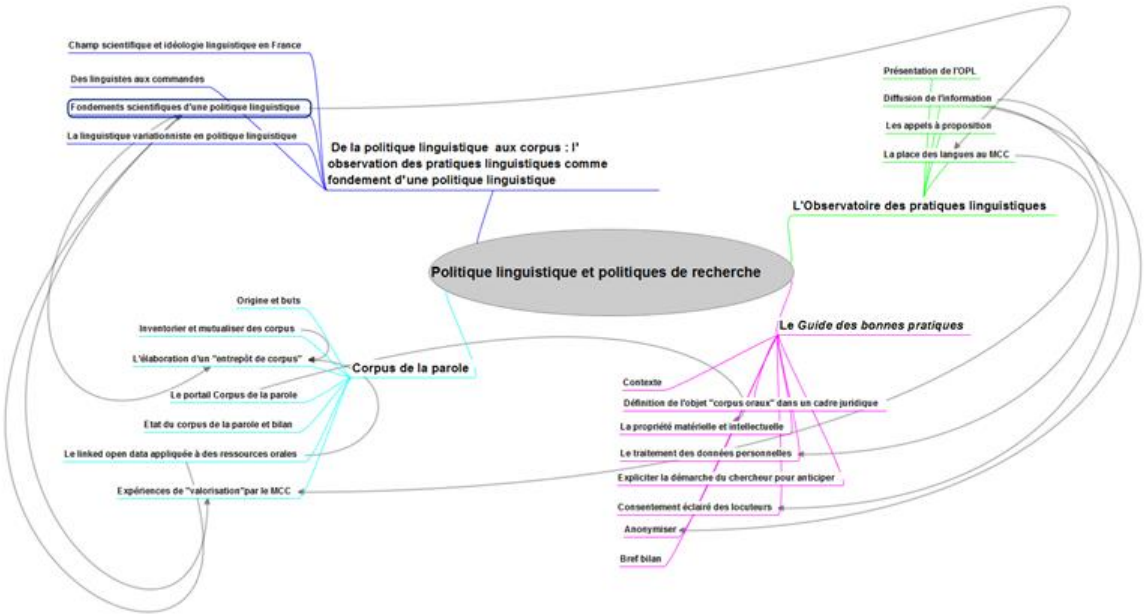
Ce travail convergera avec le chantier ouvert à la Délégation Générale à la Langue Française et aux Langues de France (Ministère de la Culture et de la Communication) pour la conception d'une version de cet outil dans le cadre d'un projet de sémantisation des données du programme « Corpus de la parole ». L'outil développé en open source sera repris et adapté pour permettre une collecte collaborative de données sonores à partir des terminaux des utilisateurs (ordinateurs, téléphones, tablettes) ou à partir d'enregistreurs numériques mis à disposition par le laboratoire. Cet outil en ligne permettra également une indexation et une description collaborative des enregistrements sonores.

En aout 2015, la réalisation d'une application Android comme premier test d'un outil collaboratif est en cours de développement par Camille Leroux.

3.8.2 Projets [retour]

Les autres projets sont décrits en conclusion.

4. Politique linguistique et politiques de recherche



4.1 De la politique linguistique aux corpus : l'observation des pratiques linguistiques comme fondement d'une politique linguistique [retour]

Articles et livre :

- 2003 Olivier Baude, Jean Sibille. L'observatoire des pratiques linguistiques. Culture et recherche, Paris : Ministère de la Culture et de la Communication, 2003, pp.7-9. https://halshs.archives-ouvertes.fr/halshs-01184590
- 2008 Olivier Baude, Michel Alessio. Les corpus de la parole : patrimoine immatériel et langues de France. Culture et recherche, Ministère de la Culture, 2008, pp.42-43. https://halshs.archives-ouvertes.fr/halshs-01184592
- 2006 Olivier Baude. Corpus oraux Un guide des bonnes pratiques. Culture et recherche, Ministère de la Culture, 2006, pp.2. https://halshs.archives-ouvertes.fr/halshs-01184593
- Olivier Baude, Jean Sibille. L'observatoire des pratiques linguistiques, suivi d'un entretien avec Pierre Ecrevé. Culture et Recherches, 2010, pp.82-83. https://halshs.archives-ouvertes.fr/halshs-01184595
- Michel Alessio, Olivier Baude. La diversité des langues. Culture et Recherches, 2010, pp.4-5. https://halshs.archives-ouvertes.fr/halshs-01184597
- Olivier Baude. L'observation des pratiques linguistiques en France. Culture et recherche, Ministère de la Culture, 1999, pp.6-8. https://halshs.archives-ouvertes.fr/halshs-01184281

Communications orales :

- Olivier Baude. « Pierre Encrevé et la réforme de l'orthographe : "le champ du linguiste ». Faire signe, pour Pierre Encrevé, Oct 2006, Paris, France. 2006. https://halshs.archives-ouvertes.fr/halshs-01165941
- 2015, Olivier Baude. Les langues de France dans le programme Corpus de la parole. Technologies pour les Langues Régionales de France, Feb 2015, Meudon, France. 2015. https://halshs.archives-ouvertes.fr/halshs-01165904

Documents :

- Site DGLFLF : http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Observation-des-pratiques-linguistiques
- http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Observation-des-pratiques-linguistiques/Langues-et-cite

4.1.1 Champ scientifique et idéologie linguistique en France

[*Linguistique variationniste*] [*retour*]

Dès la fin de ma thèse mon parcours de recherche s'est orienté vers une expérience singulière au cœur des institutions de politique linguistique en France. En effet, quelques mois après ma soutenance, j'ai intégré en 1998, en tant que vacataire, puis consultant (avec le statut de collaborateur extérieur, chargé d'études) la *Délégation Générale à la Langue Française.* La DGLF, qui dépendait du Premier Ministre de sa création en 1989 jusqu'à 1996, a été rattachée ensuite au Ministère de la Culture et de la Communication. Elle est devenue en 2001 la *Délégation Générale à la Langue Française et aux Langues de France* (DGLFLF).

La DGLFLF est chargée « *d'animer et de coordonner la politique linguistique du Gouvernement et d'orienter son évolution dans un sens favorable au maintien de la cohésion sociale et à la prise en compte de la diversité de notre société*[114] ».

Mon rôle dans cette structure a été d'assurer le lien avec la recherche en linguistique au sein de la Mission *Langues de France et Observation des pratiques linguistiques*. L'Observatoire des pratiques linguistiques a pour objectif de créer un lien entre les travaux et données de la recherche sur les pratiques linguistiques et les acteurs du monde politique et institutionnel qui sont en attente d'informations dans ce domaine. L'Observatoire a été créé en 1999 et il s'appuie sur un conseil scientifique. J'en assume la direction. Il développe et soutient des projets ou des programmes de recherche dans le cadre d'appels à proposition thématiques ou de partenariats avec le CNRS, les universités et d'autres institutions. Les terrains et les pratiques linguistiques concernent le français sous toutes ses formes, mais aussi les langues en contact avec lui, notamment les langues de France (langues "régionales", langues non territoriales et/ou langues issues des différentes vagues de migration et la LSF).

[114] http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/La-DGLFLF/Qui-sommes-nous

Avant de détailler, dans les chapitres suivants, certaines de mes activités et de les mettre en relation avec ma recherche, je préciserai la relation toute particulière qui existe, en France, entre une sociolinguistique variationniste et les activités d'un organisme dédié à la politique linguistique, une relation établie pour l'essentiel grâce au travail de Pierre Encrevé qui, s'il a beaucoup œuvré pour le développement de la sociolinguistique en France, a aussi joué un rôle de premier plan dans la politique linguistique en France depuis le milieu des années 1980.

C'est d'ailleurs, et il convient de le souligner, un positionnement parfaitement atypique d'un chercheur qui croise au plus haut niveau son activité scientifique et la « mise en société » de certains aspects théoriques, le concept de variation étant central dans cette démarche.

De fait, P. Encrevé, dès 1976, conclut ainsi sa présentation de *Sociolinguistique* de W. Labov :

> "*Sans doute une partie de la distance de l'approche linguistique de Labov et celle de Chomsky (tout aussi engagé que lui dans la lutte politique, et dans le même camp) provient-elle de ce que le premier estime, à la différence du second, que combat scientifique et combat politique sont directement liés et se renforcent.* " (Encrevé 1976:35[115])

On peut interroger cette conception de l'engagement politique et de la posture du scientifique sous différents angles, mais il reste comme caractéristique première que, dans le cas de W. Labov et de P. Encrevé – et de fait chez N. Chomsky qui en tire la conclusion inverse –, cette conception est déterminée par la définition même de l'objet de la linguistique : la langue, non exempte de sa nature sociale.

Pour comprendre l'impact de cette position à partir des actions concrètes de P. Encrevé, un rappel historique est nécessaire.

En France, l'histoire de la politique linguistique est depuis longtemps liée à une idéologie très forte qui consiste à « (…) ériger la langue française en quasi-religion d'Etat et l'unilinguisme national en principe fondateur de l'unité et de l'indivisibilité de la nation elle-même » (Encrevé 2002 :123)[116].

Je ne développerai pas cette histoire qui commence notamment par :

- la décision de l'Assemblée nationale en 1790, sur proposition du député Bouchette, "*de faire publier les décrets de l'Assemblée dans tous les idiomes qu'on parle dans les différentes parties de la France*"[117],
- le *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française* (1794) de l'Abbé Grégoire

[115] LABOV, W., & ENCREVE, P. (1976). *Sociolinguistique*.
[116] Cf. P. Encrevé, 1992, « La langue de la République », *Pouvoirs*, 100, Janvier, p.123-136.
[117] Cf. F. Brunot, 1967, *Histoire de la langue française*, Armand Colin, T.9, 1, p. 25.

- et le discours de Barère qui déclarait « *la langue d'un peuple libre doit être une et la même pour tous » et « dans une république une et indivisible la langue doit être une* »[118].

pour me consacrer à l'histoire contemporaine de la politique de la langue.

Ainsi, deux siècles plus tard, les initiatives politiques ne sont pas sans lien avec cette idéologie :

- En 1966, le président Georges Pompidou crée le « Haut Comité pour la défense et l'expansion de la langue française » ainsi que le Secrétariat permanent du langage et de l'audiovisuel. Le Haut comité pour la défense et l'expansion de la langue française deviendra en 1973 le « Haut Comité de la langue française ».
- En 1992, le Parlement adoptera l'alinéa 1er de l'article 2 de la Constitution qui stipule que : « La langue de la République est le français ».
- De même en 1994, la loi relative à l'emploi de la langue française (dite loi Toubon) donne un cadre constitutionnel et législatif à cette conception.

Le raccourci paraîtra un peu brutal mais force est de constater que cette politique linguistique, bien que très éloignée du champ scientifique (les linguistes ne sont pas à la manœuvre ni même réellement consultés), est construite sur une idéologie qui n'est pas incompatible avec la structuration de ce dernier.

J'en donnerai deux exemples :

- La défense de la langue est en France déterminée par le symbole d'une norme unificatrice. Ainsi, les langues régionales et les patois disparaissent et la dialectologie se développe rapidement en marge du champ de la linguistique institutionnelle[119] ;
- De même la linguistique, fortement dominée symboliquement par l'étude de la forme écrite de la langue, comporte au sein même de son organisation interne strictement dépendante de son origine universitaire les fondements de la description (pour ne pas dire prescription) d'une grammaire normative[120].

Une analyse du champ de la linguistique en France comme celle entreprise par J.-Cl. Chevalier, P. Encrevé et G. Bergounioux, montre que l'idéologie linguistique en France est à la fois très éloignée des théories scientifiques en linguistique tout en étant dans la logique de la structuration du champ. Depuis plus d'une vingtaine d'année, les initiatives dans le

[118] Cf. M. de Certeau, 1975, Dominique Julia, Jacques Revel, *Une politique de la langue,* Paris, Gallimard,.
119 Cf. B. Laks 2003, "Les grandes enquêtes phonologiques en France", in *la prononciation du français dans sa variation, La tribune internationale des langues vivantes* n°33.
[120] Cf. G. Bergounioux 1992, *Les enquêtes de terrain en France*, "Enquêtes, corpus, témoins", Langue française n°93, ed Larousse,Paris. "

domaine de la langue en France se sont intensifiées, une grande partie d'entre elles allant à l'encontre de cette idéologie.

Il s'agit d'une rupture que nous pouvons décrire à partir de deux grandes catégories : la reconnaissance de la diversité linguistique et l'aménagement de la langue française.

Reconnaissance de la diversité linguistique :

- 1988 : référendum sur la Nouvelle-Calédonie reconnaissant les droits culturels des Kanak et prévoyant l'organisation d'un enseignement des langues kanak à l'école maternelle ;
- 1995 : loi sur le statut de la Polynésie dont l'alinéa 1er de l'article 115 énonce : « Le français étant langue officielle, la langue tahitienne et les autres langues polynésiennes peuvent être utilisées »;
- 1998 : accords de Nouméa reconnaissant les langues kanak comme langues de culture et d'enseignement et prévoyant la formation d'enseignants et la création d'une Académie des langues kanak ;
- 1999 : signature par la France de la Charte européenne des langues régionales ou minoritaires, qui a nécessité la mise au point d'une liste des langues de France concernées devant être annexée à la Charte. Cette liste comprend 75 langues de France, l'indivisibilité constitutionnelle de la République impliquant en effet de compter les langues des DOM et TOM au nombre des langues régionales, ce qui bouleverse la vision courante. Cette signature a été suivie de la décision du Conseil constitutionnel interdisant la ratification par le Parlement de l'adhésion à la Charte ;
- 2001 : la DGLF, placée depuis 1996 auprès du Ministère de la Culture, devient la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF) ;
- 2002 : la loi relative à la Corse reconnaît l'enseignement de la langue corse « dans le cadre de l'horaire normal des écoles maternelles et élémentaires de Corse » (art. 7), malgré une très vive polémique nationale ;
- 2015 : Tentative de ratification de la Charte.

Aménagement de la langue française :

- 1989 : création par décret du Conseil Supérieur de la langue française (CSLF), présidé par le Premier Ministre, et de la Délégation Générale à la Langue Française (DGLF), placée auprès du Premier Ministre ;
- 1990 : rapport au Premier Ministre du CSLF sur les « Rectifications de l'orthographe », publié au *Journal Officiel* (Documents administratifs) et approuvé par l'Académie française ;

- 1998 : circulaire Jospin invitant les administrations à la féminisation des noms de métiers, titres, grades et fonctions ;
- 2001 : les ministres de la Culture et de la Réforme de l'Etat créent le Comité d'Orientation pour la Simplification du Langage Administratif (COSLA) qu'ils président en personne et qui se donne notamment pour tâche de réécrire les grands formulaires administratifs nationaux.
- 2015 : Création de l'Agence de la langue française

On peut constater, d'une part, un emballement des initiatives en matière de politique de la langue : une quinzaine en à peine plus de 15 ans, et, d'autre part, une rupture très nette avec l'idéologie linguistique prévalente. Or, il se trouve que ces initiatives révèlent le rôle primordial de linguistes, P. Encrevé en tête, et qu'elles ont toutes comme caractéristique de bouleverser les relations entre théorie linguistique et politique de la langue.

4.1.2 Des linguistes aux commandes [retour]

La première caractéristique de cette transformation est la place donnée à la science dans la politique de la langue :

- Un renforcement de l'avis des experts : rapport sur l'orthographe, rapport sur la féminisation, rapport sur les langues régionales etc.
- La création de l'*Observatoire des pratiques linguistiques* : lieu de rencontre entre "les scientifiques et les décideurs". La méthodologie de la linguistique n'est pas directement le lieu de la politique linguistique. On peut néanmoins signaler la création par P. Encrevé de *l'Observatoire des pratiques linguistiques* au sein de la DGLFLF en 1999 avec pour mission " *d'étudier les pratiques linguistiques en France ainsi que les modalités et les effets du contact entre les langues, afin d'apporter des informations utiles pour l'élaboration des politiques sociales, éducatives et culturelles en permettant de prendre en compte les expériences linguistiques des individus et des groupes.*" *Le champ de l'observation est celui des pratiques linguistiques actuelles. Il s'agit donc de travaux sociolinguistiques sur l'usage actuel du français et des langues utilisées en France. Les données rassemblées doivent provenir principalement d'enquêtes, d'entretiens ainsi que de corpus attestés, constitués de productions réelles, dans une situation donnée, de locuteurs nettement identifiés. Elles doivent porter sur la description sociolinguistique des usages et des variations du français standard et non standard, des langues régionales, des langues de l'immigration et des langues dites « sans territoire »*[121].
- un rôle d'acteurs directs :

[121] DGLFLF 2000, Rapport au parlement.

- En 1988, le Premier Ministre nomme à son cabinet, chose singulière, un linguiste qu'il charge entre autres de la langue française.
- Dans la foulée, les organismes créés (Conseil supérieur de la langue française et Délégation Générale à la Langue Française) passent sous la responsabilité de linguistes.
- Lors de sa création, le COSLA sera vice-présidé par un linguiste.

On pourrait concevoir que ce rôle d'expert et d'acteur dévolu aux linguistes relève d'une science appliquée. Ce n'est pas vraiment le cas est c'est dû, principalement, à la théorie linguistique mobilisée, telle qu'elle a été proposée par P. Encrevé : il ne s'agit pas d'une sociologie de la langue mais d'une sociolinguistique qui veut être, bien plus qu'un domaine de la linguistique générale, la linguistique de la variation.

4.1.3. Fondements scientifiques d'une politique (linguistique) de la variation [retour]

La nature sociale de la langue [Linguistique variationniste]

La première caractéristique de la sociolinguistique est la réinterprétation de la dichotomie langue/parole (Labov 1973 – Encrevé 1976, 1977) :

> " *Car, pour refuser ce triple déploiement de la dichotomie langue/parole, il suffit de rétablir au premier plan ce qui crève les yeux, que le social s'il est uni est aussi divisé, qu'il est le champ de contradictions et d'affrontements et que la langue (comme système, structure, machine) est partie prenante et partie prise de ces divisions.*"[122]

Ainsi la linguistique variationniste de P. Encrevé repose sur la reconnaissance qu'il y a, dans la langue, de la variation sociale et qu'"*un système social peut user de la variation langagière pour organiser la société notamment par l'usage de la domination, de légitimation/ délégitimation, de classement, enfin l'ensemble des distinctions que le social opère en ce domaine*"[123].

Dans cette conception, les usages sociaux de la variation sont un des lieux où s'agence le rapport entre la langue et le social et il faut y reconnaître un lieu d'enjeux sociaux. Une

[122] Encrevé, 1976, "Labov, linguistique, sociolinguistique", in Labov *Sociolinguistique*, éditions de Minuit, Paris, p 12.
[123] Cf. Encrevé, 2006, "Variations" entretien avec G. Bergounioux, in *Faire-Signe colloque international en hommage à Pierre Encrevé*, Orléans, PUO, p 68.

conception de la langue en tant que structurant et structurée par les relations sociales est décisoire pour l'implication de la théorie dans la politique de la langue. Il n'est pas étonnant qu'elle fournisse matière à des initiatives politiques.

Une nouvelle pratique linguistique

La deuxième caractéristique de la linguistique variationniste est de proposer une nouvelle pratique du linguiste fondée sur l'observation de données socialement situées.

> "*Une analyse du langage hors contexte subsistera sans aucun doute en tant que domaine autonome (…) mais désormais la théorie linguistique ne pourra pas plus dédaigner le comportement social des sujets parlants que la chimie ne peut ignorer les propriétés observables des éléments.*" (Labov 1976 :350-351[124])

> *"Pour atteindre cet objectif il a fallu tout d'abord doter la linguistique d'une méthodologie de l'enquête construite sur une armature sociologique : nécessité d'une analyse de la communauté sociale et des conditions de l'observation.*" (Encrevé 1988 :14)[125].

Pour ces deux premiers points, c'est dans un dialogue continu avec la sociologie de P. Bourdieu, et notamment avec ses travaux sur les biens symboliques, que P. Encrevé a conforté sa théorie linguistique.

> "*Il m'a paru, en effet, que le point de vue sociolinguistique exigeait l'intégration d'une problématique sociologique systématique. Cette sociologie joue d'abord comme vigilance épistémologique, et intervient comme telle au cœur même de l'élaboration théorique proprement linguistique, là où on ne soupçonnerait pas sa présence. Elle joue aussi, cela va de soi, dans la construction de l'enquête comme sociologie de la population enquêtée et sociologie de la relation sociale d'enquête ; comme sociologie de la langue, enfin, elle est active à tous les niveaux de l'analyse des données. Sur tous ces points, la problématique de Bourdieu m'apparait la plus apte à permettre non pas d'appliquer à la linguistique une sociologie toute faite, mais de construire dans la pratique sociolinguistique une connexion étroite entre linguistique et sociologie – où chacune des deux disciplines s'arme de l'autre pour se poser à soi-*

[124] LABOV, W., & ENCREVE, P. (1976). *Sociolinguistique*.
[125] ENCREVE, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*.

> *même des questions nouvelles, sans crainte exagérée du brouillage des frontières.*" (Encrevé 1988 :15)[126]

Il en ressort notamment la mise en œuvre de concepts centraux : *capital*, *habitus*, *légitimité*, *marché linguistique*.

Une linguistique générale

La troisième caractéristique de la linguistique variationniste est de réfuter la réduction de la variation en linguistique à une variation sociale ou même à une co-variation : "la variation est inhérente aux langues humaines, indépendamment même d'une variation sociale."[127] Ainsi la langue commune est à concevoir comme un ensemble d'invariance et de variations.

Il y a ici un double enjeu : premièrement une prise en compte du social qui ne se réduise pas à une covariation :

> " *La finalité de l'approche "variationniste" en linguistique ne consiste pas, comme le supposent parfois encore les linguistes classiques, à décrire les variations linguistiques pour les rapporter à leur distribution sociale.*" (Encrevé 1988 :14).

Deuxièmement, il s'agit de situer la linguistique variationniste comme proche du programme de la linguistique générative :

> *« La linguistique, en effet, s'est définie de se donner comme objet la langue et non la parole et le discours, la compétence et non les performances, et ce choix, qui a rendu possible la définition actuelle de la linguistique comme science cognitive, science des grammaires intériorisées que se construit l'enfant dès sa mise au monde, n'est mise en cause par aucun linguiste, pas plus Labov que nous-même, quitte, bien sûr à déplacer le lieu de la frontière entre langue et parole, pour construire des grammaires qui rendent compte de leur utilisation effective dans la vie ordinaire des locuteurs-auditeurs que nous sommes tous : des grammaires qui intègrent l'hétérogénéité, la variation inhérente au système, qu'elles doivent d'abord au fait que l'enfant construit sa ou ses grammaires dans une communauté linguistique toujours, et dès le premier jour, socialement différenciée, clivée, divisée*[128].

Variations, hétérogénéité et grammaire du récepteur

[126] Ibid.
[127] idem
[128] Encrevé 2003 "langue et domination", Intervention au CSE, hommage à Pierre Bourdieu.

Enfin, quatrième caractéristique de la linguistique variationniste : le rôle donné à la réception dans le circuit de la parole.

> "*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. 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 multitudes de paroles dont l'ensemble hétérogène arrivera à former la langue de la communauté*[129]".

Cette primauté donnée à la réception est centrale dans la reconnaissance de l'hétérogénéité bien évidemment, mais elle l'est également dans la théorie phonologique et dans le concept de grammaire d'auditeur. Elle a aussi un impact fort sur le développement d'une théorie linguistique de la réception, par la suite particulièrement importante dans le développement de la sociopragmatique.

4.1.4 La linguistique variationniste en politique linguistique

[Linguistique variationniste] *[retour]*

Cette présentation de l'appareillage théorique de la linguistique variationniste est par trop sommaire, néanmoins, les grandes lignes du bouleversement des initiatives politiques en découlent.

Cette conception de la langue est le fondement même d'une politique linguistique de la variation qui s'est traduite par :

- La reconnaissance de l'usage de la variation langagière par des systèmes sociaux pour structurer la société (*capital, légitimité, habitus, marché linguistique*.)
- La reconnaissance de la variation comme inhérente à la langue y compris en termes de capacité cognitive du sujet (le monolinguisme n'existe pas, les variations sont inhérentes aux usages de la langue et cela concerne également les variétés de langues et le multilinguisme).
- L'interprétation de cette double reconnaissance en termes de droits linguistiques du citoyen et notamment dans le cadre de l'article XI de la Déclaration des droits de l'homme et du citoyen: « *La libre communication des pensées et des opinions est un des droits les plus précieux de l'homme : tout citoyen peut donc parler, écrire, imprimer librement, sauf à répondre de l'abus de cette liberté dans les cas déterminés par la loi* ».
- La reconnaissance des langues de France.

[129] Encrevé 1977, "Présentation : linguistique et sociolinguistique", Linguistique et sociolinguistique, *Langue française* n°34, Larousse, p. 6.

- La modification de loi Toubon.
- Le choix de la féminisation des noms de métiers et des titres par les usages.
- Les propositions de variantes orthographiques et le respect de la diversité des pratiques linguistiques en matière de représentation alphabétique des mots.

Bien d'autres éléments relevant du champ de la politique linguistique en France (Société des agrégés, Association de défense du français, Académie française, attitude du corps enseignant etc.), toutefois la politique linguistique en France apparait comme fortement liée à un cadre théorique selon lequel compte en premier *la variation linguistique qui ne refoule pas la nature sociale de la langue*.

C'est dans ce cadre de lien entre recherche en linguistique et politique linguistique qu'il faut analyser les actions entreprises durant une quinzaine d'années. *[science et politique]*

4.2 L'Observatoire des pratiques linguistiques

4.2.1 Présentation de l'OPL *[retour]*

Dans les documents de présentation à partir de 1999, L'Observatoire des pratiques linguistiques est défini de la façon suivante :

> *« Installé à la DGLFLF et doté d'un conseil scientifique, l'observatoire a pour mission de recenser et de rendre disponibles tous les savoirs relatifs à la situation linguistique en France. C'est un comité d'experts, qui ne fait pas de recherches lui-même, mais lance des appels d'offres thématiques et subventionne les travaux de laboratoires universitaires et autres unités de recherche. Le champ de l'observation est celui de la sociolinguistique et concerne les pratiques actuelles, qu'il s'agisse du français ou des autres langues parlées sur le territoire national. Les données rassemblées proviennent d'enquêtes de terrain, et rendent compte des expériences langagières réelles des individus et des groupes. Elles portent aussi bien sur l'hétérogénéité des usages (variations géographiques ou sociales), que sur les questions de contact de langues, de transmission ou d'acquisition ; sur les modalités du plurilinguisme comme sur les évolutions en cours (féminisation, déplacement des normes, effets des supports de l'écrit sur la langue…).*
>
> *Le rôle de l'observatoire est aussi de favoriser la collaboration et l'organisation en réseau des équipes et centres de recherche qui travaillent sur les pratiques linguistiques sur l'ensemble du territoire et dans les pays francophones. Les résultats des recherches et l'ensemble des données*

recueillies sont intégrés dans une banque de données gérée et actualisée par la DGLFLF. »

En tant que directeur scientifique de l'Observatoire, j'ai eu à coordonner depuis quinze ans l'ensemble de ses travaux, de sa définition à sa programmation en passant par les partenariats et les grands projets[130].

Les deux principaux objectifs de l'observatoire sont :

- O1 : Recenser, développer et rendre disponibles les savoirs relatifs à la situation linguistique en France.
- O2 : Faire mieux connaitre un patrimoine linguistique commun, constitué par l'ensemble des langues et des variétés linguistiques parlées en France.

En 2015, trois missions répondent à ces prescriptions :

- La diffusion des informations recueillies auprès des spécialistes, des responsables de politiques publiques et du large public.
- Le soutien à des travaux d'étude et de recherche, la coordination et l'organisation en réseaux de ces travaux.
- La constitution, la conservation et la diffusion de données sur les pratiques linguistiques dont, en premier lieu, les corpus de « pratiques linguistiques ». Ces corpus constituent un outil de travail pour la recherche, mais acquièrent également, avec le temps, un caractère patrimonial.

Le fonctionnement de l'Observatoire repose sur un conseil scientifique présidé par P. Encrevé et dont les membres permanents sont : le président du CS, B. Laks, O. Baude, le Délégué général à la langue française et aux langues de France.

Le CS expertise les projets scientifiques qui sont soumis à la DGLFLF, notamment dans le cadre des appels à propositions. Il émet également un avis sur les thèmes des publications de l'OPL, *Langues et Cité* et *les Cahiers* de *l'Observatoire des pratiques linguistiques*. Enfin le conseil scientifique est consulté sur les différents dossiers qui sont de son ressort (projets soumis spontanément, partenariats avec le CNRS, Programme Corpus de la parole, projets des autres missions...). Sa composition a été différente selon les années (représentant de l'AUF, directeurs des fédérations de recherche en linguistique, représentant du MCC et des chercheurs et/ou spécialistes d'un domaine). L'OPL travaille étroitement avec la mission Langues de France. Son budget est d'environ 100 à 150 000 euros par an. Tous les projets aidés financièrement sont soumis à un avis du conseil scientifique. En 15 ans, environ 150

[130] Notons tout l'intérêt de la thèse de sciences politiques soutenue par Frédérique Niel en 2007 qui présente une étude approfondie de l'Observatoire des pratiques linguistiques. NIEL, F., & LABORIER, P. (2009). *Les vicissitudes de l'État linguiste ou Comment les langues minoritaires deviennent l'objet d'une politique sociale linguistique: contribution à une sociologie historique du capital informationnel d'état*.

projets ont bénéficié d'une aide financière (la moyenne des aides se situe entre 3 et 12 000 euros, ces montants sont loin d'approcher ceux de l'ANR ou d'autres appels à projets de recherche). Au sein de la DGLFLF, l'OPL dépend du chargé de mission "Langues de France et Observation des pratiques linguistiques"

Deux exemples de composition du Conseil scientifique :

CS 11 mai 2000 (Expertise Appel à projets) :

Mme Magnant, déléguée générale à la langue française
M. Ladousse, conseiller au cabinet de la ministre de la culture et de la communication
M. Encrevé, directeur d'études à l'E.H.E.S.S
M. Cerquiglini, vice-président du Conseil supérieur de la langue française, directeur de l'INaLF
M. Chevalier, professeur émérite, membre du Conseil supérieur de la langue française
M. Laks, professeur à l'université Paris X-Nanterre
Mme Blanche-Benveniste, professeur à l'université d'Aix-en-Provence
Mme Auzanneau, maître de conférences à l'université Paris V
M. Bergounioux, professeur à l'université d'Orléans
M. Boyer, professeur à l'université Montpellier III
M. Rézeau, directeur de recherches à l'INaLF
Mme Delamotte-Legrand, professeur à l'Université de Rouen
Mme Gadet, professeur à l'université Paris X-Nanterre,
Mme Fattier, professeur à l'université Paris X-Nanterre,
M. Rabaud, DDAT
Mme Rouot, mission de la recherche et de la technologie
Mme Fuchs, ministère de la recherche, excusée
M. Salles Lousteau, inspecteur général de l'éducation nationale
M. Le Divennah, FAS, excusé
M. Mahieux, Délégation interministérielle à la ville
M. Héran, directeur de l'INED
M. Jean, chef de mission à la DGLF
M. Alessio, chargé de mission à la DGLF
M. Catillon, chargé de mission à la DGLF
M. Baude, chargé d'études à la DGLF

CS 22 octobre 2008 :

M. Xavier North,
M. Pierre Encrevé,
M. Michel Alessio,
M. Jean Sibille,
M. Olivier Baude,
M. Benoit Habert (directeur adjoint d'ADONIS, directeur de l'ILF,)
Mme Stéphane Robert (directrice de la fédération TUL),

Mme Naila Louise-Rose (secrétaire générale des fédérations ILF et TUL).

4.2.2. Diffusion de l'information : [retour]

Langues et Cité
Sous le contrôle du conseil scientifique, j'ai créé en1999 un bulletin d'une douzaine de pages en moyenne, *Langues et Cité*, afin de diffuser les résultats des recherches financées par l'Observatoire. J'ai été rédacteur en chef du bulletin pendant treize ans avant de confier cette responsabilité à Valelia Muni Toke.
Ce bulletin a eu très rapidement un franc succès. Tiré selon les numéros entre 1500 et 4000 exemplaires, il est également diffusé sur le site de la DGLFLF. Le public visé est celui des personnes intéressées (politiques, documentalistes, travailleurs sociaux etc.) par une mise à jour des informations relatives aux études sociolinguistiques et aux langues pratiquées dans la vie quotidienne par les citoyens. Le rôle du rédacteur en chef est de rédiger l'éditorial et de sélectionner les articles qui composent les numéros thématiques. Ce second aspect a nécessité un dialogue permanent avec de nombreux chercheurs, notamment afin d'obtenir des textes accessibles au lectorat ciblé. Cette expérience contribue à une autre forme de réflexion sur nombre de notions linguistiques dans leur confrontation au terrain.

Langues et Cité, le bulletin de l'Observatoire des pratiques linguistiques

http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Observation-des-pratiques-linguistiques/Langues-et-cite

- n° 26 : les langues Kanak de Nouvelle-Calédonie, décembre 2014
- n° 25 : Le francique (Platt Lorrain) mars 2014,
- n° 24 : Féminin, masculin la langue et le genre, octobre 2013
- n° 23 : Le berbère, aout 2013
- n° 22 : Le corse, U corsu, décembre 2012
- n° 21 : Le catalan – novembre 2012
- n° 20 : Nouvelles technologies, nouveaux usages ? - octobre 2011
- n° 19 : Parler (avec) plusieurs langues : l'alternance codique - octobre 2011

- n° 18 : Le francoprovençal - janvier 2011
- n° 17 : Le breton - juillet 2010
- n° 16 : Langues en contact - mars 2010
- n° 15 : L'arabe en France - octobre 2009
- n° 14 : Des hommes, des langues, des pratiques - juillet 2009
- n° 13 : Plurilinguisme et migrations - novembre 2008
- n° 12 : Langues d'ici, langues d'ailleurs - juillet 2008
- n° 11 : L'arménien en France - février 2008
- n° 10 : L'occitan - décembre 2007
- n° 9 : La langue (r)romani - juin 2007
- n° 8 : Des langues dans la cité - décembre 2006
- n° 7 : Les rectifications orthographiques - septembre 2006
- n° 6 : Corpus de la parole - mai 2006
- n° 5 : Les créoles à base française - octobre 2005
- n° 4 : La langue des signes française - novembre 2004
- n° 3 : Les langues en Guyane - mai 2004
- n° 2 : Les pratiques langagières des jeunes - septembre 2003
- n° 1 : Observer les pratiques linguistiques : pour quelles politiques ? - octobre 2002

La lecture des sommaires et des éditoriaux témoigne de la répartition entre des numéros consacrés à la présentation et à la mise en perspective des langues de France et des numéros consacrés à un thème lié à l'observation des pratiques linguistiques. On retrouve les deux fondements d'une « politique de la variante » : la diversité des pratiques et la construction d'une politique sur des savoirs scientifiques à l'épreuve d'études empiriques.

Les Cahiers de l'Observatoire

Les *Cahiers de l'Observatoire* sont destinés à publier des études linguistiques dans leur intégralité (environ 150 pages). Trois numéros ont paru, un numéro sur la LSF est sous presse.

Les Cahiers de l'Observatoire

http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Observation-des-pratiques-linguistiques/Langues-et-cite/(offset)/20

- 2012, *Langues de France, langues en danger* : aménagement et rôle des linguistes, Cahiers de l'observatoire des pratiques linguistiques n°3,
- 2008, *Migration et plurilinguisme en France*, Cahiers de l'observatoire des pratiques linguistiques n°2,
- 2006, *Les rectifications orthographiques de 1990 : analyses des pratiques réelles,* Cahiers de l'observatoire des pratiques linguistiques n°1.

Le Rapport au parlement

La DGLFLF rédige chaque année un *Rapport au parlement sur l'emploi de la langue française* de 150 pages. C'est un outil d'évaluation de la politique linguistique en France. Il contient systématiquement un résumé des actions de l'Observatoire qui peuvent ainsi être connues et discutées par les parlementaires. A notre connaissance, il s'agit d'une expérience unique de description systématique d'une politique linguistique.

Rapport au parlement sur l'emploi de la langue française 2014

http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/La-DGLFLF/Nos-priorites/Rapport-au-parlement-sur-l-emploi-de-la-langue-francaise-2014

Les travaux de l'Observatoire des pratiques linguistiques

Créé en 1999 au sein de la Délégation générale à la langue française, l'Observatoire des pratiques linguistiques a pour objectif de recenser, de développer et de rendre disponibles les savoirs relatifs à la situation linguistique en France, afin de fournir des éléments d'information utiles à l'élaboration des politiques culturelles, éducatives ou sociales. Il a également pour but de faire mieux connaitre un patrimoine linguistique commun, constitué par l'ensemble des langues et des variétés linguistiques parlées en France, qui concourent à la diversité culturelle de notre pays.

L'activité de l'Observatoire s'organise autour de quatre axes :

>> le soutien à des travaux d'étude et de recherche, la coordination et l'organisation en réseaux de ces travaux ;

>> la diffusion des informations recueillies auprès des spécialistes, des responsables de politiques publiques et d'un large public ;

>> l'organisation en réseau et la collaboration des équipes et centres de recherche qui travaillent sur les pratiques linguistiques en France et dans les pays francophones ;

>> la participation de la DGLFLF aux projets structurant la recherche sur le français et les langues de France.

Depuis sa création, l'Observatoire a procédé à 8 appels à propositions thématiques (en 1999, 2000, 2001, 2005, 2008, 2010, 2012, 2013). L'appel à projets 2013 concernait les pratiques langagières en langues de France. La qualité des projets présentés a conduit la DGLFLF à élaborer une liste complémentaire de projets retenus au titre du budget 2014 : 5 projets de recherche présentés par différentes universités ou laboratoires du CNRS ont été aidés en plus des 9 projets retenus en 2013. La première phase d'activité de l'Observatoire a consisté à mobiliser les chercheurs et à favoriser l'émergence de réseaux. La seconde phase consiste à créer des espaces nouveaux de diffusion de l'information et d'échange avec les décideurs, les acteurs sociaux, les acteurs culturels soucieux de disposer de données scientifiques. Pour cela, un bulletin, *Langues et cité*, a été créé. Pour 2014, deux numéros ont été programmés : le n°25 Le francique (platt lorrain) et le n°26, à paraitre, sur les langues kanakes.

En 2006, l'Observatoire avait inauguré une collection de publications intitulée Les *Cahiers de l'Observatoire des pratiques linguistiques* : le n°1, intitulé Les rectifications orthographiques de 1990 : analyses des pratiques

réelles en France et dans la francophonie, est paru en 2006. La collection a été relancée avec le n°2 : Migrations et plurilinguisme en France, paru en septembre 2008 à l'occasion des États généraux du multilinguisme. Le n°3, Langues de France, langues en danger : aménagement et rôle des linguistes, paru fin 2012, est constitué par les actes de Journées d'étude organisées en partenariat avec l'université de Lyon II en 2010.
Depuis 2004, la DGLFLF entretient un partenariat avec les fédérations de recherche en linguistique du CNRS (l'Institut de la Langue Française et la fédération Typologie et Universaux Linguistiques). Ce partenariat se concrétise par le soutien à des initiatives structurantes comme le Congrès Mondial de Linguistique Française ou comme le développement du programme Corpus de la parole. Ce programme est dédié à la conservation, la constitution, la mise à disposition et la valorisation de corpus oraux (sous la forme de collection de documents sonores enregistrés à des fins d'analyses linguistiques). Ces corpus constituent un outil de travail pour la recherche, mais acquièrent également, avec le temps, un caractère patrimonial. Ce programme, lancé dans le cadre du plan de numérisation du Ministère de la Culture et de la Communication, permet de constituer et de numériser une collection de corpus oraux en français et en langues de France, mise à la disposition du public sur le site http://corpusdelaparole.huma-num.fr/, ouvert en février 2008.
En 2009 cette priorité s'est traduite par la signature avec le CNRS d'une convention en vue du développement de ce programme qui se poursuit en 2014. Son objectif est non seulement le développement d'une base de données patrimoniales sur l'oral, mais aussi le développement d'outils de traitement automatique des langues et d'ingénierie linguistique. Le département des archives sonores de la Bibliothèque nationale de France (BnF) est un partenaire privilégié dans ces projets.
Ces différents programmes ont accéléré la participation de la DGLFLF aux projets récents consacrés à la recherche en linguistique :

>> soutien à la création en 2012 et au développement (2013-2014) d'une Unité Mixte de Recherche (universités d'Orléans et de Tours, CNRS et BnF) sur les corpus oraux ;

>> relations avec la Très Grande Infrastructure de Recherche en Humanités Numériques françaises (Huma-Num) qui, conformément à la feuille de route Horizon 2020, est le porteur de la participation française à l'ERIC Dariah créé en 2014 et anime deux consortiums consacrés à la linguistique ;

>> participation au comité d'orientation de l'Equipex Ortolang dédié aux ressources et outils du français et des langues de France. Cet équipement d'excellence créé dans le cadre des investissements d'avenir a pour but de proposer, pour l'ensemble de la communauté de recherche française en linguistique, une infrastructure offrant un réservoir de ressources (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement. Il a pour mission de permettre, au travers d'une mutualisation des ressources, le développement de la recherche sur l'analyse, la modélisation et le traitement automatique du français afin de la hisser au meilleur niveau international ; de faciliter l'usage et le transfert des ressources et des outils des laboratoires publics vers les partenaires industriels ; de valoriser le français et les langues de France par un partage des connaissances accumulées par les laboratoires publics.

4.2.3. Les appels à propositions. [retour]

La seconde mission de l'OPL est de soutenir des travaux de recherche en sociolinguistique. Depuis 1999, l'Observatoire a procédé à huit appels à propositions thématiques :

- 1999*, Observation du contact linguistique dans une situation géographique et sociale précise-1*
- 2000, *Observation du contact linguistique dans une situation géographique et sociale précise-2*
- 2001, *Transmission familiale et acquisition non didactique des langues*
- 2005, *La LSF*
- 2008, *Valorisation et usages de corpus oraux*

- 2010, *Observation, description, analyse de l'alternance codique, français, langues de France et autres langues en contact sur le territoire français*
- 2012, *Numérique et textualité : observation, description et analyse des pratiques contemporaines*
- 2013. *Observation des pratiques linguistiques en langues de France*

Au total, ce sont 85 projets qui ont été financés lors de ces appels à propositions. En tant que directeur scientifique, j'ai pris en charge la rédaction du texte de l'appel, la gestion des propositions, la procédure d'évaluation à partir d'une grille validée par le conseil scientifique, la gestion des réponses après validation de la DGLFLF sur avis du conseil scientifique et enfin le suivi du dossier au sein des services gestionnaires de la DGLFLF, du CNRS et des universités.

Cette expérience des appels à propositions de la DGLFLF appelle plusieurs remarques :

1. Le thème du premier appel était très général et il devait permettre de faire un état des lieux de la recherche en sociolinguistique en France. De fait, à partir des propositions, on peut analyser la constitution du domaine. Une analyse reste à mener mais on constate que les appels suivants ont été plus ciblés, ce qui démontre une relative faiblesse du champ et une impossibilité de piloter une recherche sur un vaste domaine ou de définir une thématique prioritaire.
2. Le nombre de propositions soumises (de 7 à 39) donne également une information sur la vitalité de la recherche en sociolinguistique.
3. La relative diversité des structures qui déposaient une proposition confirme quelle difficulté rencontre une institution comme la DGLFLF pour trouver un partenaire académique correspondant directement au champ de ses préoccupations en termes de politiques sociales éducatives et culturelles sur les langues.

Aussi, un retour sur les appels à propositions a décidé d'une orientation pour construire un partenariat avec les deux fédérations de recherche et l'élaboration de programmes sur des thématiques à long terme.

4.2.4. La place des langues au sein du MCC [retour]

L'évaluation de l'action de l'Observatoire peut également se faire en prenant en compte l'évolution de la place des langues en général et des pratiques linguistiques en particulier dans les différentes initiatives et projets du Ministère de la Culture et de la Communication.

S'il est vrai que la diversité linguistique et la variation – partie prise, partie prenante du monde social –, n'est pas encore véritablement reconnue et si l'on constate un manque cruel dans ce domaine de recherche, il n'en reste pas moins que l'Observatoire, de par ses actions directement sur le terrain de la recherche, a permis à la politique linguistique conduite par la DGLFLF d'être intégrée systématiquement aux travaux du Ministère de la

Culture. Ainsi, par exemple, et nous reviendrons sur ce point, depuis 2005, le plan de numérisation du Ministère inclut un chapitre concernant les pratiques linguistiques en français et en langues de France. D'une manière plus générale, les travaux de l'Observatoire figurent depuis 2007 dans l'accord-cadre du CNRS et du MCC.

4.3 Le Guide des bonnes pratiques

Articles et livre :	
	○ 2006, *Corpus oraux, Guide des bonnes pratiques*, https://halshs.archives-ouvertes.fr/halshs-00355472 ○ 2007, Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux, https://halshs.archives-ouvertes.fr/halshs-01163043 ○ 2008, Le droit de la parole, https://halshs.archives-ouvertes.fr/halshs-01162543 ○ 2010, Spoken Corpora Good Practice Guide 2006, https://halshs.archives-ouvertes.fr/halshs-01165893 ○ 2010, Corpus oraux, Guide des bonnes pratiques 2006. Version allemande, https://halshs.archives-ouvertes.fr/halshs-01165896 ○ 2015, Recherche des indices permettant une identification : l'anonymisation des transcriptions du corpus ESLO, https://hal.archives-ouvertes.fr/hal-01174647
Communications orales :	
	○ 2005, Transcrire : les bonnes pratiques des linguistes, https://halshs.archives-ouvertes.fr/halshs-01162548 ○ 2006, Constitution et exploitation d'un grand corpus de "données situées" Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), https://halshs.archives-ouvertes.fr/halshs-01165954 ○ 2007, Mutualiser des corpus oraux, aspects juridiques et déontologiques, https://halshs.archives-ouvertes.fr/halshs-01166001 ○ 2007, Constituer et exploiter un corpus d'interactions- aspects juridiques et éthiques, https://halshs.archives-ouvertes.fr/halshs-01166002 ○ 2013, Les linguistes de l'oral et les données numériques, https://halshs.archives-ouvertes.fr/halshs-01165993 ○ 2014, Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO, https://halshs.archives-ouvertes.fr/halshs-01165957 ○ 2014, Pratiques de linguiste : du juridique à l'éthique, de la collecte à la diffusion, https://halshs.archives-ouvertes.fr/halshs-01165956 ○ 2014, Pratiques numériques en SHS, « (Bonnes) pratiques du document sonore : l'exemple du programme Corpus de la parole, https://halshs.archives-ouvertes.fr/halshs-01165991
Documents :	
	○ Site ESLO, aspects juridique, http://eslo.huma-num.fr/index.php/pagemethodologie?id=69 ○ Formulaire de consentement ESLO2 : http://www.nakala.fr/data/11280/df5c9365 ○ Charte Ortolang : https://www.ortolang.fr/#/information/policy

4.3.1. Contexte [retour]

Dès le début des années 2000, premières années d'existence de l'Observatoire des pratiques linguistiques, celui-ci a été confronté aux problèmes suivants :

- Le manque de corpus disponibles en matière de pratiques linguistiques. Dans la plupart des cas, les chercheurs qui avaient participé à un projet où qui détenaient un corpus exprimaient leurs difficultés à diffuser des données pour lesquelles ils n'étaient pas assurés de maitriser les aspects juridiques.

- Le besoin de consignes à formaliser pour la collecte et la diffusion des corpus fournis par les responsables de projets subventionnés par la DGLFLF.

- La reprise de données anciennes dans le cadre du plan de numérisation du Ministère de la Culture et des opérations d'archivage et de diffusion de l'information scientifique et technique.

- L'ambition de déployer une politique française et européenne pour les « données libres », notamment autour de l'open data.

Ces préoccupations étaient partagées par de nombreux linguistes, pour la plupart récemment sensibilisés aux questions éthiques et juridiques. Ainsi l'ASILA- RTP 14 de 2003 (http://www.loria.fr/projets/asila/doc/CRlaBresse.pdf) qui a réuni plusieurs jours durant des linguistes spécialisés dans les corpus, a mis en place un groupe de travail sur les aspects juridiques. Celui-ci a continué dans la structure qui a poursuivi ce travail : l'EPML 50 en 2004. De son côté, l'équipe CLAPI du laboratoire ICAR travaillait sur les aspects juridiques des corpus et bases de données de la linguistique de l'interaction.

Ce constat offrait l'opportunité, pour la DGLFLF, de mener un programme de recherche appliquée. En effet, d'une part, depuis la naissance des Archives de la parole, le Ministère de la Culture joue un rôle important dans la gestion des données orales issues de la recherche. D'autre part la DGLFLF recherchait un programme afin de fédérer les actions et projets de recherche dans le cadre de son partenariat avec les fédérations de recherche en linguistique. J'ai ainsi été conduit à présenter au conseil scientifique, en tant que directeur scientifique de l'Observatoire, le projet d'un programme sur les aspects juridiques des corpus oraux en 2003 puis en 2004.

L'originalité de la démarche de la DGLFLF a été de constituer un groupe de travail autour d'un laboratoire d'études juridique (le CECOJI) sous la direction d'Isabelle de Lamberterie, avec des linguistes, des conservateurs et des informaticiens spécialisés en gestion de corpus. Ce groupe de travail, que j'ai piloté, s'est réuni à partir de la fin 2003 :

« *Le programme général sur les corpus oraux est le programme prioritaire de l'Observatoire des pratiques linguistiques de la DGLFLF. Les propositions faites au conseil scientifique de l'Observatoire qui se tiendra le 7 décembre 2004, confirment cette volonté de la DGLFLF de concevoir le soutien à la constitution, la conservation et la diffusion des corpus oraux comme une action de politique linguistique. La première action est la constitution d'un groupe de travail constitué de membres de la DGLFLF, de la BNF, de l'INA, de la DMF, de juristes du CNRS et de linguistes (représentants des deux fédérations du CNRS, Institut de la Langue Française et Typologie et Universaux Linguistiques, plus des responsables de projets scientifiques) afin d'aider les équipes de recherche à normaliser les pratiques de recueil et d'exploitation de corpus au regard de la législation en prenant en compte l'ensemble des contraintes liées à la recherche.*
L'objectif est de rédiger un « guide de bonnes pratiques *» servant d'aide aux chercheurs et accompagnant la méthodologie d'une communauté scientifique :*

- *Il est fondamental de concevoir le travail du groupe comme une construction collective des chercheurs en collaboration avec les juristes.*

- *Pour rédiger ce guide il convient tout d'abord de réaliser une typologie des situations d'usage des corpus (recueil, exploitation, valorisation, diffusion).*
- *La qualification juridique nécessaire pour le guide doit s'appuyer sur une grille de typologie des situations. »*

(Baude, Note interne DGLFLF, 2004)

Il faudra compter quinze mois avant la publication de l'ouvrage collectif : *Corpus oraux, Guide des bonnes pratiques 2006*. Nous pouvons noter, dès l'origine, la spécificité de ce travail et qui en conditionnera le succès. Le *Guide* a bénéficié de la collaboration des juristes autour d'Isabelle de Lamberterie à partir d'une démarche d'accompagnement de la réflexion d'une communauté scientifique et non d'une simple expertise des contraintes légales. Ce point est primordial et le *Guide* n'aurait pas été aussi utile à la communauté s'il n'avait pas bénéficié d'une élaboration par des juristes à l'écoute de pratiques, de méthodologies et d'objectifs scientifiques.

Cet accompagnement est défini dès les premières pages du Guide :

« Pour remplir sa mission, ce groupe de travail s'est donné pour objectifs notamment de :

- recenser les pratiques actuelles et définir en priorité les contraintes méthodologiques et théoriques liées à la recherche ;

- diffuser une synthèse sur la législation existante ;

- établir des recommandations ;

- et, le cas échéant, en cas de vide ou de flou, formuler des propositions pour l'élaboration de normes et règles juridiques (notamment européennes).

Il fallait pour cela tout d'abord :

- recenser les domaines juridiques concernés ;

- identifier et quantifier les risques ;

- repérer les réponses existantes ;

et ensuite construire ces réponses sous la forme d'une série de recommandations de bonnes pratiques (juridiques et éthiques). (GBP 20)

Cette méthode qui part des pratiques de terrain, en les accompagnant par un travail d'expertise, pour permettre la définition de « bonnes pratiques » prises en charge par une communauté a été le fil conducteur. En ce sens, on peut rapprocher cette expérience des objectifs développés par la suite dans le cadre des infrastructures de recherche en humanités numériques qui ont vu le jour quelques années plus tard. La structuration de la TGIR Huma-Num à partir de consortiums disciplinaires et de services développés par des spécialistes participe de ce même mouvement qui concilie les pratiques des chercheurs et les contraintes de politique scientifique.

Pour cette réflexion à partir des aspects juridiques, l'apport des juristes est directement visible :

La méthode à laquelle s'est rallié le groupe de travail se caractérise par les traits suivants :

- la conviction qu'il ne faut pas laisser croire qu'il existe des réponses toutes faites à tout type de situation ;

- la volonté de ne pas « brider » les chercheurs (en interdisant certaines pratiques par exemple) ;

- le respect de la méthodologie du chercheur et des contraintes liées à l'observation (les chercheurs souhaitent enregistrer des situations sans que les contraintes, notamment techniques et juridiques, les modifient).

- la nécessité d'élaborer et de rédiger ce guide en mettant en commun les compétences requises aux différentes étapes (linguistes, juristes, conservateurs) ;

- l'affichage d'une démarche fondée sur le respect de la loi et de l'éthique ;

- la nécessité de fournir à travers ce Guide un outil d'expertise des risques (repérage, mais aussi évaluation). (GBP 21)

Le recensement des pratiques des linguistes de corpus nécessita une collaboration avec des équipes témoins pratiquant le recueil de données orales ou audio-visuelles afin de définir une typologie des situations françaises et internationales. Afin de compléter cette approche, le groupe de travail a rédigé un questionnaire que le directeur scientifique de l'Observatoire eut en charge de distribuer auprès d la communauté.

A partir de ces différents résultats, une première liste des questions les plus fréquentes posées par les chercheurs a été dressée :

1. *Quelles autorisations dois-je faire signer aux locuteurs que j'enregistre pour pouvoir ensuite exploiter ce corpus et pouvoir :*
   a. le citer dans un travail universitaire ;
   b. le citer dans un article publié dans une revue scientifique ;
   c. le citer dans un ouvrage à diffusion commerciale ;
   d. le mettre à disposition sur un site ;
   e. le diffuser sur CD.
   Ces différents types d'exploitation sont-ils soumis aux mêmes règles ?
2. *J'ai fait un enregistrement de personnes que je connais bien.*
   a. A quelles conditions puis-je l'exploiter ? (exploiter est pris au sens de la question 1)
   b. Peuvent-elles revenir sur leur autorisation ?
3. *Lorsque j'enregistre des enfants,*
   a. qui peut donner son consentement ?
   b. lorsque l'enfant sera majeur peut-il revenir sur ce consentement ?
   c. si l'enregistrement a lieu dans le cadre scolaire, faut-il des autorisations particulières ?
4. *Dans le cadre d'un travail au sein d'un laboratoire,*
   a. Qui est considéré comme l'auteur du corpus ?
   b. Quel(s) droit(s) ce travail donne-t-il au chercheur ?
5. *Qui est considéré comme « responsable » de la diffusion et du traitement d'un corpus ?*
6. *Si je masque les noms propres de personnes, cela suffit-il pour que je puisse utiliser librement une transcription ?*
7. *Sous quelles conditions puis-je archiver mon corpus sous la forme de fichiers informatiques ?*
8. *Si les personnes que j'ai enregistrées (dans les médias ou en privé) sont décédées, ai-je une liberté d'exploitation de ces enregistrements ?*
9. *Je découvre dans une armoire des enregistrements. Je voudrais pouvoir les exploiter. Je n'ai plus la trace de qui a enregistré ou qui a été enregistré.*
   a. Puis-je me servir de ces documents ?
   b. Quelles précautions (quelles garanties) dois-je prendre ?
10. *J'enregistre une émission à la radio (ou à la télévision).*
    a. Puis-je utiliser librement la transcription ?
    b. Puis-je utiliser la version sonore ?
    c. Du point de vue des autorisations, y a-t-il une différence entre émissions des radios publiques et des radios privées ?

d. Y a-t-il une différence entre enregistrer des personnalités connues et enregistrer des « anonymes » (personnes qui témoignent, s'expriment en libre antenne, auditeurs qui posent des questions, etc.) ?
e. les droits d'exploitation sont-ils différents si j'achète une cassette, un dévédé ou un cédé de l'émission ou si j'enregistre moi-même l'émission lorsqu'elle est diffusée ?

11. *J'aimerais constituer un corpus de données authentiques. Quelles sont les précautions que je dois prendre ?*

Cette liste de questions positionnait l'ouvrage à partir d'exemples très concrets. Il n'en reste pas moins que si les pratiques exposées sont bien réelles et fréquentes, les questions sont rarement explicites dans les projets. Inexistantes dans le projet ESLO1, on n'en trouve également nulle trace dans le numéro de *Langue française* dirigé par G. Bergounioux en 1992 et qui représente un élément important pour saisir le degré de réflexivité du champ en ce domaine : *Enquêtes, corpus et témoins en France, hier et aujourd'hui*[131].

« Prenant en compte les cadres juridiques existant en France (et plus généralement dans un certain nombre de points en Europe), ce guide s'appuie sur les questionnements des chercheurs qui ont participé à son élaboration. Ceux-ci ont cherché à comprendre les fondements des règles juridiques applicables et les enjeux liés à leur respect et à leur mise en œuvre. C'est donc une *vision dynamique de la régulation juridique* qui sert de trame à ce guide, à travers la démarche que suivent les chercheurs. Les auteurs du *Guide*, eux-mêmes impliqués sur les terrains de recherche dont il est question ici, ont eu le souci de proposer des pratiques et usages respectueux des droits existants. Pour cela, la démarche du chercheur doit consister à connaitre l'existence de ces droits et des contraintes qui en découlent. Il s'agira ensuite de tirer les conséquences de ces contraintes tant dans la phase du recueil des données que dans celle de leur valorisation. » GBP : 22

Le contexte général du projet repose donc sur une ambigüité entre une bonne volonté manifeste et une difficulté à construire une réflexion qui soit intégrée aux cadres théoriques de la linguistique de corpus. Les aspects juridiques deviennent un enjeu majeur mais en périphérie d'une discipline. Cela prouve, s'il en était besoin, que la linguistique reste dominée par le principe de disparition du locuteur comme agent du monde social et par celui de « l'exception scientifique » qui autorise le chercheur à ne pas approfondir ses interrogations sur les questions d'éthique.

De fait les problèmes sont réels :

- Les possibilités de diffusion rapide et en masse créent une nouvelle situation qui implique directement le chercheur, le directeur de labo et les institutions.

- La synchronisation son/texte dévoile un nouvel objet qui donne toute sa place au signal, à la voix et donc au locuteur (légitimant du même coup les données sociologiques identifiantes utilisées par la sociolinguistique, l'ethnolinguistique et d'autres secteurs disciplinaires.

[131] BERGOUNIOUX, G. (1992). *Enquêtes, corpus et témoins en France, hier et aujourd'hui*.

- Dans le passage de l'oral au graphico-visuel, de nombreuses opérations de catégorisation sont effectuées, soit quant aux formes linguistiques, segmentées visuellement en unités (Blanche-Benveniste & Jeanjean, 1987 ; Mondada, 2000), soit quant à l'identité des locuteurs eux-mêmes (Mondada, 2003) avec des risques de stéréotypisation (Jefferson 1996) et de stigmatisation des locuteurs et de leurs façons de parler, d'autant plus que le corpus bénéficie d'une diffusion plus large.

On montrera qu'il est possible d'envisager une autre approche du droit, de la linguistique et des politiques de recherche à partir d'un travail collectif entre différentes disciplines et différents métiers, à partir d'une approche réflexive qui permet d'expliciter les démarches.
Le début des années 2000 a vu évoluer une situation jusqu'alors figée et on peut relever à présent des éléments de contexte – techniques, scientifiques et politiques – plus favorables.

Contextes techniques :
Les nouvelles technologies de traitement des corpus oraux offrent une opportunité :

- Données sonores numérisées,
- Manipulation et diffusion aisées,
- Transcriptions synchronisées,
- Outils d'aide à la transcription,
- Avancées importantes en reconnaissance et synthèse de la parole,
- Développement d'outils favorisant l'interopérabilité.

Contextes scientifiques :

- Programme des fédérations de recherche en linguistique du CNRS,
- Initiative ASILA-EPML50 et projets de collaboration d'équipes,
- Projets de constitution de grand corpus (C-oral-Rom, PFC…),
- Appel d'offre ANR "corpus" appelé à devenir récurrent,
- Création des centres de ressources numériques (CRDO),
- Développement d'initiatives de normalisation et interopérabilité (Olac…),
- Modification du champ de la linguistique (l'oral, les langues de petite diffusion deviennent un objet d'étude légitime)
- etc.

Contextes politiques :

- *« Les organismes publics doivent avoir le souci constant de faire bénéficier au mieux la collectivité nationale des fruits de leurs travaux... ».*
« La politique de la recherche et du développement technologique vise à l'accroissement des connaissances, à la valorisation des résultats de la recherche, à la diffusion de l'information scientifique et technique et à la promotion du français comme langue scientifique. »
Art 5 de la Loi n°82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France, aujourd'hui art. L 111-1 du code de la recherche. *JO* du 16-07-1982, p. 2273 et ss.

- Le 22 octobre 2003, à Berlin, la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) ont signé la Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales, dont l'objectif est de promouvoir Internet « comme instrument fonctionnel au service d'une base de connaissance globale de la pensée humaine ».

- Les programmes de numérisation patrimoniale

En 2001 à Lund, en Suède, un groupe de représentants nationaux des États membres de l'Union européenne, intéressés par les problèmes de numérisation, a élaboré un texte qui prône notamment : la mise en place de standards d'interopérabilité ; la diffusion de bonnes pratiques dont la gestion des droits de propriété intellectuelle ; l'organisation de centres de compétences sur la numérisation dont les professionnels de l'information ont la responsabilité. En France, au sein du plan de numérisation du Ministère de la Culture, les langues deviennent pour la première fois en 2007 l'une des thématiques reconnues.

Ces contextes sont favorables à une prise en charge de « bonnes pratiques » par une communauté scientifique qui doit assumer de l'ensemble des aspects, suivant une présentation des productions respectant les pratiques des chercheurs et leurs sources et se conformant à un objectif technique et scientifique : l'interopérabilité et la diffusion des données.

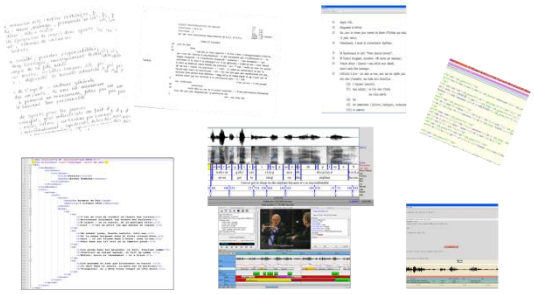
4.3.2. Définition de l'objet « corpus oraux » dans un cadre juridique [retour]

La forme des corpus oraux est relativement complexe. Dans la majorité des cas, les corpus oraux sont constitués :

- d'enregistrements (analogiques ou numériques) qui, en cas de supports analogiques, ont une durée de vie très courte avec une perte de qualité lors des migrations,

- de données contextuelles sur les locuteurs et la situation d'enquête qui peuvent être en partie des données personnelles (nom propre, profession, adresse, lieu…),

- de transcriptions (sous la forme de fichiers indépendants ou permettant une synchronisation sur le signal ; transcription phonétique, orthographique, multilinéaire, etc.),



- d'annotations "secondaires" (informations sur les conditions de production des énoncés, précisions sur les phénomènes sonores tels que les rires et les bruits),
- d'annotations enrichies (étiquetage morphologique, syntaxique, annotations prosodiques, pragmatiques…),
- de documentation.

Ainsi, d'une façon très schématique, la réponse aux questions juridiques consiste à :

- définir le statut juridique de l'objet "corpus" (quelles sont les conditions d'élaboration et de quoi est-il composé ?),
- procéder à la gestion contractuelle des droits des personnes concernées,
- définir les responsabilités de ceux qui vont intervenir dans la vie du corpus (créateurs, hébergeurs, diffuseurs…).

L'analyse de l'objet « corpus oraux » et des pratiques et « situations d'usage » des chercheurs permet de repérer trois domaines juridiques principaux :

- La propriété matérielle et intellectuelle,
- Le traitement des données personnelles,
- La responsabilité des hébergeurs et des diffuseurs.

4.3.3 La propriété matérielle et intellectuelle [ex : ESLO] [retour]

- Afin de gérer les droits liés à la propriété matérielle et intellectuelle, il convient de définir si un corpus oral recèle une création protégée par le droit d'auteur.

  – Soit le corpus est constitué d'œuvre du domaine public. « *Le domaine public recouvre non seulement les idées de liberté d'accès et de gratuité d'utilisation des données, mais aussi la possibilité pour chacun de les exploiter.* »

  Dans ce cas il est libre de droits : liberté d'accès, gratuité d'utilisation, liberté d'exploitation, de fait ou après 70 ans…

  – Soit il est protégé par le droit d'auteur à condition qu'il réponde aux trois conditions suivantes :

  - qu'il corresponde à l'exigence d'une **activité créatrice :** un travail de compilation d'informations n'est pas protégé en soi.

- qu'il ait une **forme définie.** Ce qui est protégé, ce n'est pas le contenu du corpus mais son enveloppe, son architecture.

- « *Enfin, la forme du corpus doit répondre à la condition d'être* ***originale****. Que signifie l'originalité d'un corpus ? L'originalité de nombreuses créations de l'ère du numérique, comme les logiciels ou les bases de données, ne peut être appréciée que d'après des critères objectifs. Il semble qu'il en soit de même des corpus oraux, ceux-ci pouvant le plus souvent être assimilés à une base de données. C'est alors, le plus souvent, le fait que le corpus soit ou non copié et révèle un minimum d'activité créative qui servira de critère pour déterminer s'il est ou non original (et non pas uniquement la prise en compte de l'empreinte de la personnalité de son auteur)* » (GBP :39).

- Une fois la question de la nature de la création résolue, c'est la question de la définition de l'auteur qui se pose.
  - L'auteur est en principe la (ou les) personne(s) physique(s) sous le nom de laquelle (ou desquelles) l'œuvre est divulguée.
  - En cas de pluralité d'auteurs, 2 cas :
    - il s'agit de co-auteurs (mêmes droits pour chacun des auteurs)
    - on est en présence d'une œuvre collective : peuvent être qualifiées d'œuvre collective celles créées « sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et sous son nom, et dans laquelle la contribution personnelle des divers auteurs se fond dans l'ensemble » Art. L. 113-2 du CPI.
  - Dans ce dernier cas, c'est la personne physique ou morale qui a pris l'initiative de l'œuvre qui dispose des droits d'auteur.
- Se pose alors la question des **droits qui concernent les corpus protégés**. Il s'agit :
  - Des droits patrimoniaux (autoriser ou interdire la reproduction ou la communication au public).
  - De la prérogative du droit moral (divulgation, retrait, paternité, respect de l'œuvre).

Précisons que *« Le chercheur auteur qui refuse de divulguer le corpus qu'il a créé est dans son droit (au titre du droit d'auteur), même si par ailleurs il peut être sanctionné administrativement pour ne pas avoir exécuté sa mission de service public qui est de communiquer les résultats de sa recherche ». (GBP :41)*

- Enfin, il faut définir si le corpus relève du domaine public. C'est une vaste question qui semble encore discutée selon les approches de la discipline mais quoi qu'il en soit, il existe une solution intermédiaire : la liberté d'accès et d'exploitation.

« Sans être dans le domaine public, ces corpus sont – de par la volonté de leurs créateurs – libres d'accès et d'utilisation. Néanmoins, si les créateurs peuvent renoncer à exercer leurs droits patrimoniaux, il ne leur est pas possible de renoncer à leur droit moral, qui reste imprescriptible ». (GBP : 39).

4.3.4. Le traitement des données personnelles [ex : ESLO] [retour]

Il est fondamental de déterminer si un corpus oral contient des données personnelles. En effet, si c'est le cas, il y a obligation de se conformer à la loi et de respecter la loi Informatique et libertés. En revanche, si les données sont anonymisées, elles sortent du champ de la loi.

Qu'est-ce qu'une donnée personnelle ? Les fiches (en l'occurrence juridiques) présentes dans le guide ont pour but d'aider le chercheur à appréhender les notions nécessaires à l'explicitation de sa démarche et à la résolution des problèmes juridiques. Ainsi, *« La loi du 6 janvier 1978 s'articule autour de la notion de donnée nominative. La Convention 108 du Conseil de l'Europe de 1981 lui préfère celle de données personnelles et la directive 95/46/CE choisit l'expression « donnée à caractère personnel* » (GBP:107)

Derrière la notion de « données personnelles » se profile la question de l'identification d'une personne. Or, « (…) *pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne* » (GBP :107).

Le cas de la voix semble encore confus entre une approche juridique qui considère la voix comme une donnée identifiante donc personnelle et certains linguistes qui s'appuient sur les recherches en reconnaissance de la parole notamment pour dire qu'il est impossible d'identifier avec certitude une personne à partir de sa voix.

La question de la qualification des données dépend de la phase de « traitement » qui est définie par l'article 2b de la directive européenne :

« *toute opération ou ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. (…)* » La longueur de la définition illustre le champ des possibilités ouvert par l'outil informatique, tout en l'étendant à tous les types de traitement. La distinction entre traitement automatisé et non automatisé n'a plus cours. Il en va de même pour la notion de fichier, lorsque le législateur

européen met sur le même plan les fichiers informatiques et manuels : il suffit que les données soient organisées suivant une structure définie.

« *Le traitement n'implique pas forcément une manipulation du fichier, un simple stockage suffit à le faire entrer dans le champ d'application. La difficulté vient une nouvelle fois de la très grande portée de la définition.* » (GBP:108)

Il faut donc anonymiser les données qui permettent une identification directe ou indirecte en prenant en compte une *probabilité suffisante de rapprochement et le fait que le récepteur constitue un élément important*.

Tout traitement des données doit avoir un responsable. Ainsi la directive européenne 95/46/CE ("Flux de données transfrontières") dans son article 2d, repris pour la refonte de la loi "informatique et libertés", donne la définition suivante :

"Le responsable d'un traitement de données à caractère personnel est, sauf désignation expresse par les dispositions législatives ou règlementaires relatives à ce traitement, la personne, l'autorité publique, le service ou l'organisme qui détermine ses finalités et ses moyens." (GBP :108)

Le responsable du traitement se doit donc de veiller à :

1. la qualité des données ;
2. à leur(s) finalité(s) ;
3. au recueil du consentement.

En outre il est est responsable de la déclaration à la CNIL : "Tout traitement de données personnelles à caractère direct ou indirect doit être déclaré à la CNIL".

Quelques années après la publication du *Guide*, nous pouvons tirer un bilan de cette première partie du travail autour de la définition des cadres juridiques.

Tout d'abord, la démarche du *Guide* est d'aider le chercheur à prendre en charge les enjeux de ses pratiques au regard des aspects juridiques. Une réflexion conjointe de linguistes et de juristes ouvre des pistes de recherches inédites sur la notion même de « données personnelles » et de « données identifiantes ». Ainsi, dans le cadre du projet ESLO, la phase d'anonymisation a ouvert un vaste chantier sur les entités nommées et le repérage automatique des données identifiantes (Eshkol, Baude, Maurel, 2008, 2011, 2015 et chapitre XXX).

Ensuite, la définition des responsabilités et de la propriété intellectuelle place le chercheur dans une démarche ou son rôle se doit d'être explicite et défini. Il est de fait primordial que le chercheur *sache ce qu'il fait*. C'est d'ailleurs le second apport structurant du *Guide*. Le dialogue entre juristes et linguistes doit s'appuyer sur l'explicitation de la démarche du

chercheur dès lors que pour définir le statut juridique d'un corpus, il faut en connaitre les conditions d'élaboration et d'exploitation.
Ainsi la prise en charge des aspects juridiques revient à :

- définir le statut du corpus en explicitant la démarche (de la collecte à la diffusion),
- anticiper les exploitations, y compris la diffusion et le changement de finalité,
- recueillir le consentement,
- traiter les données personnelles.

4.3.5. Expliciter la démarche du chercheur pour anticiper [ESLO] [retour]

L'originalité du travail mené afin d'aboutir au *Guide des bonnes pratiques* repose sur la part de réflexivité de la démarche d'observation qui est le point de départ de la constitution du corpus. Ainsi il n'y a pas de réponse juridique sans un réel travail d'explicitation de leur démarche par les chercheurs eux-mêmes. Les aspects juridiques découlent d'une description fine du type de données, des techniques d'enquêtes (et notamment le mode d'approche), le lieu de la collecte, le dispositif d'enregistrement et enfin l'exploitation, de la transcription à la diffusion.

Seule cette explicitation de la démarche permet de générer de bonnes pratiques en anticipant le recueil de consentement, l'anonymisation, les traitements, la conservation, la diffusion et les changements de finalité.

Comment expliciter la démarche de création et d'exploitation de corpus oraux ? Celle-ci peut être décrite à partir des grandes phases qui scandent un travail sur corpus :

Les enregistrements qui constituent les données primaires de l'enquête linguistique sont loin de former un objet uniforme. Ainsi, un conte enregistré sur une bande magnétique lors d'une cérémonie traditionnelle sur la place d'un village est un objet scientifique et patrimonial fort différent de l'enregistrement numérique d'un texte lu par un « informateur rémunéré » dans les locaux d'un laboratoire universitaire, des réponses à un questionnaire enregistrées sur minidisque par un chercheur au domicile de la personne interrogée ou bien encore d'une conversation spontanée non sollicitée par les chercheurs, se déroulant dans un café et filmée par une ou plusieurs caméras.

Il convient donc, dans un premier temps, d'identifier les éléments qui caractérisent les données récoltées en situation :

-le *type de données* qui constitue le corpus et leurs supports (d'enregistrement, mais aussi de stockage pour exploitation, et de conservation),

- les *différentes techniques* employées par les chercheurs pour récolter les données,
- la définition des *participants* et de leur rôle,
- la catégorisation des *lieux* de la collecte.

GBP :43

Expliciter les objectifs du projet

Un corpus est tout d'abord défini par les buts du projet de recherche :

- Analyses (cadres théoriques et diffusion des analyse : rapports, colloques, articles…)
- Diffusion des données primaires
- Diffusion des données enrichies (avec des outils de fouilles : concordancier, langage de requête, etc.)
- Exploitation (ingénierie, didactique…)
- Conservation (stockage, stockage sécurisé, archivage pérenne)

Ce point semble évident mais une rapide analyse des projets fondées sur des corpus oraux permet de constater que nombre d'objectifs sont implicites soit parce que les cadres théoriques conditionnent ces objectifs, soit parce que ceux-ci sont passés sous silence par les responsables du projet. Enfin, l'hypothèse d'un changement de finalités n'est quasiment jamais évoquée.

Type de données :

Un corpus oral est composé d'enregistrements qui peuvent être audio ou vidéo et qui sont ensuite décrits à l'aide de métadonnées puis transcrits et annotés. La plupart du temps, ces données sont ensuite structurées en BDD. Si le type d'enregistrement dépend principalement de la technologie, les descriptions et annotations sont fortement dépendantes des effets de normalisation et de standardisation. Sur ce point, les linguistes français sont organisés à travers les consortiums de la TGIR Huma-Num et leur participation à des projets internationaux comme CLARIN. La TEI reste en 2015 l'initiative qui a l'impact le plus important sur le type de données produites par les linguistes avec le groupe ISO TC37 SC4. Ces aspects de normalisation et de standardisation des données sont développés dans le chapitre sur les cadres théoriques. Il convient de rappeler l'objectif de faire converger, dans le guide, les questions de codage et les questions juridiques tout autant que théoriques :

> *« Ainsi, les principes qui doivent guider le choix d'une technologie plutôt que d'une autre pour l'annotation peuvent être résumés en quatre questions :*
> – *Cette technologie permet-elle de coder de manière explicite toutes les annotations ?*
> – *Cette technologie présente-t-elle un caractère propriétaire ou une limite légale qui empêcheraient de partager les annotations avec d'autres (formats propriétaires, techniques basées sur des brevets etc.) ?*
> – *Cette technologie est-elle acceptée par la communauté avec laquelle l'échange des données est envisagé ?*
> – *Cette technologie a-t-elle fait l'objet d'une normalisation ? »* GBP : 47

Techniques d'enquêtes :

Un corpus oral est défini par les objets qui le composent et par les techniques d'enquête utilisées. En linguistique, celles-ci sont très diversifiées :

- Enregistrement en laboratoire
- Questionnaire
- Entretien
- Recueil de contes, chants…
- Récit de vie
- Activités dans leur contexte ordinaire
- Reprise d'enregistrements

Toutefois les données enregistrées ne sont pas des données préexistantes et recueillies, elles sont toujours le produit de la situation d'enquête (Cameron 1992[132]).
Ainsi, les techniques d'enquêtes produisent (ou valident) une catégorisation nominale des participants – informateurs, témoins, sujets, cobayes, natifs, enquêtés, collaborateurs, observés, participants, acteurs sociaux… Cette catégorisation résulte de statuts mais aussi de rôles construits par la technique d'enquête : comme par exemple le « témoin » VS « l'observateur-participant ».

Les participants à l'enquête et aux activités enregistrées sont catégorisables de différentes manières, qui toutes éclairent de façon spécifique ce qu'ils font et ce qu'ils disent (Sacks, 1972). Ainsi les participants à une situation d'enregistrement peuvent-ils être à la fois considérés comme des enquêtés (si l'on rapporte la situation au fait qu'elle est un objet d'enquête) et comme des acteurs sociaux – dont la caractérisation précise dépend du contexte, de l'activité, des formes d'engagement et de participation, impliquant à la fois l'histoire sociale des personnes et l'accomplissement local de leur rôle, mais aussi de leur identité durant la rencontre. Selon la manière dont les chercheurs eux-mêmes traitent ces multiples catégories, différentes conséquences peuvent apparaitre à la fois pour l'objet de l'enquête et pour l'évaluation du caractère plus ou moins sensible de l'activité.

CATEGORIES DE PARTICIPANTS

La terminologie très variée utilisée dans la littérature pour définir les catégories de participants à une enquête révèle des implications éthiques et théoriques diverses (Cameron et al., 1991) (…)

Ces choix terminologiques sont le plus souvent le produit de considérations théoriques et politiques qui révèlent le type de relations préexistantes, construites, ou développées entre l'enquêté et l'enquêteur.

Si nous ne pouvons développer ici les enjeux de ces considérations théoriques, il est néanmoins important de repérer les marques d'une relation particulière qui fonde différentes réalisations de la

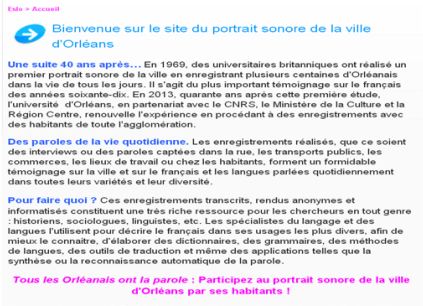
[132] CAMERON, D. (1992). *Researching language: issues of power and method*.

paire enquêté/enquêteur, impliquant différents droits et obligations selon les caractéristiques de cette relation (Sacks, 1972).

Deux éléments définissent notamment cette relation : la proximité/distance des participants et les rôles en action et en situation. GBP :52

Les techniques d'enquête produisent également la construction d'une relation sociale à partir du mode d'approche aux implications éthiques et personnelles déterminantes :

- Exemple de mode d'approche en milieu urbain (ESLO2)


Esslo > Accueil

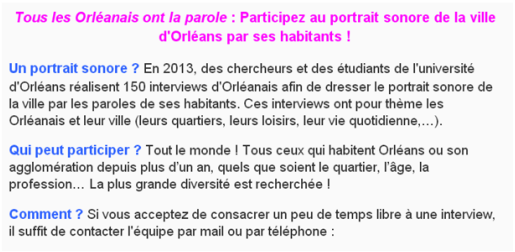
Bienvenue sur le site du portrait sonore de la ville d'Orléans

Une suite 40 ans après... En 1969, des universitaires britanniques ont réalisé un premier portrait sonore de la ville en enregistrant plusieurs centaines d'Orléanais dans la vie de tous les jours. Il s'agit du plus important témoignage sur le français des années soixante-dix. En 2013, quarante ans après cette première étude, l'université d'Orléans, en partenariat avec le CNRS, le Ministère de la Culture et la Région Centre, renouvelle l'expérience en procédant à des enregistrements avec des habitants de toute l'agglomération.

Des paroles de la vie quotidienne. Les enregistrements réalisés, que ce soient des interviews ou des paroles captées dans la rue, les transports publics, les commerces, les lieux de travail ou chez les habitants, forment un formidable témoignage sur la ville et sur le français et les langues parlées quotidiennement dans toutes leurs variétés et leur diversité.

Pour faire quoi ? Ces enregistrements transcrits, rendus anonymes et informatisés constituent une très riche ressource pour les chercheurs en tout genre : historiens, sociologues, linguistes, etc. Les spécialistes du langage et des langues l'utilisent pour décrire le français dans ses usages les plus divers, afin de mieux le connaître, d'élaborer des dictionnaires, des grammaires, des méthodes de langues, des outils de traduction et même des applications telles que la synthèse ou la reconnaissance automatique de la parole.

***Tous les Orléanais ont la parole* : Participez au portrait sonore de la ville d'Orléans par ses habitants !**


***Tous les Orléanais ont la parole* : Participez au portrait sonore de la ville d'Orléans par ses habitants !**

Un portrait sonore ? En 2013, des chercheurs et des étudiants de l'université d'Orléans réalisent 150 interviews d'Orléanais afin de dresser le portrait sonore de la ville par les paroles de ses habitants. Ces interviews ont pour thème les Orléanais et leur ville (leurs quartiers, leurs loisirs, leur vie quotidienne,...).

Qui peut participer ? Tout le monde ! Tous ceux qui habitent Orléans ou son agglomération depuis plus d'un an, quels que soient le quartier, l'âge, la profession... La plus grande diversité est recherchée !

Comment ? Si vous acceptez de consacrer un peu de temps libre à une interview, il suffit de contacter l'équipe par mail ou par téléphone :

- Exemples de *fieldwork* (travail de terrain). Celui-ci implique des relations de confiance du chercheur et d'établissement d'un contact personnel lors d'un micro-trottoir où les participants sont choisis plus ou moins aléatoirement, en raison de leur présence sur le lieu d'intervention.

Décrire la situation : le lieu

L'information sur le lieu de la collecte conditionne des éléments de réponses juridiques particuliers par ses caractéristiques propres et le rôle qu'il joue dans la situation d'enquête :

> « *Ainsi on peut tout d'abord différencier les lieux publics, au sein desquels l'activité scientifique d'enregistrement audio-vidéo ne requiert pas d'autorisation autre que celle de la personne enregistrée, et les lieux privés, soumis à l'autorisation préalable du propriétaire/responsable qui est distincte du recueil du consentement de l'enquêté.*
>
> *Le lieu peut également être défini selon la relation que les participants établissent. S'agit-il d'un lieu où la présence de la personne enregistrée est du fait de l'enquêteur (laboratoire, salle d'enregistrement…) ou est-ce celui-ci qui se déplace sur le terrain et investit donc l'espace propre de l'enquêté ?*
>
> *Enfin, le lieu d'enregistrement peut être intégré aux données (caractéristiques audio ou visuelles présentes dans les données) ou ne*

relever que d'une information éventuellement présente parmi les métadonnées ». (GBP 54)

Traitement des données

Le traitement des données linguistiques et sociologiques est une phase supplémentaire à expliciter dans la démarche du chercheur. Il serait illusoire de séparer l'activité de « terrain » des traitements qui sont anticipés et réalisés simultanément ou a posteriori. Selon Mondada (Mondada 2004) nous sommes systématiquement face à la « *transformation du terrain en un espace domestiqué conforme aux ordres des phénomènes recherchés et des analyses qu'ils subiront* ». Pour cela le chercheur utilise des techniques qui

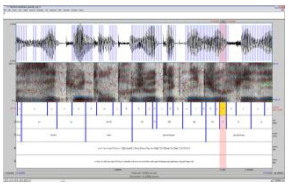
- "tamisent" les données, en éliminant les bruits provenant du terrain,
- rendent les données "compatibles" avec les analyses, les calculs, la formalisation dont elles feront l'objet,
- assurent la "comparabilité" des données.

Dans le cas de corpus linguistiques, on peut établir une liste des phases de traitement des données :

- Enrichissement (annotations)
- "Appauvrissement" (destruction, caviardage pour anonymisation etc.)
- Codage
- Catalogage (catégorisation)
- Formatage
- Standardisation/Normalisation
- Structuration (Base de données)

Un exemple de transcription (avec effets de traitements des données, catégorisation et identification du locuteur en tête) :

ESLO2 Transcription « linguiste de l'oral »



XG14:

Ben euh gé-géotechnique géologie géophysique voilà j'ai touché un petit peu à tout ce qui a rapport avec euh le terroir le sous-sol l'étude géologique tout ce qui est ben interprétation du du faciès géologique de surface et en profondeur à donc à cette époque-là ben je cumulais euh trois trois études entre géotechnique géophysique euh géologie euh bah pour faire son son trou un petit peu parce que ce type de métier il est lié au prix du baril de pétrole

ESLO 2 : transcription journalistique
Sébastien Martin, 28 ans, géologue :

> Une expérience ordinaire…
> Voilà j'ai touché un peu à tout ce qui a rapport avec le terroir, le sous-sol, l'étude géologique. Tout ce qui est interprétation du faciès géologique de surface et en profondeur.
>
> Donc à cette époque-là, je cumulais trois études entre géotechnique, géophysique, géologie. Pour faire son trou, car ce type de métier est lié au prix du baril de pétrole et à l'époque le prix du baril de pétrole était très faible, nos professeurs nous disaient : « *tant que le baril fera pas vingt-cinq dollars vous ne trouverez pas de boulot* » et donc j'ai joué un petit peu les mercenaires en faisant monter les enchères.
> -6-

Ces exemples démontrent les effets de la catégorisation et des autres processus induits par la démarche du chercheur sur l'ensemble de la chaine de constitution, conservation et exploitation de corpus oraux.

Le travail réflexif sur la démarche présenté succinctement permet d'élaborer des pratiques qui tiennent compte des aspects juridiques et éthiques lors des deux phases du process véritablement concernées : le recueil de consentement et l'anonymisation.

4.3.6 Consentement éclairé des locuteurs *[ex : ESLO]* *[retour]*

Le recueil de consentement des locuteurs est un exemple significatif du rapport des méthodes en linguistique de corpus avec la « nature sociale » de la langue. Si on fait une brève typologie des pratiques des chercheurs, nous trouverons une certaine hétérogénéité :

- projets sans recueil de consentement, soit parce que l'activité scientifique est considérée comme supérieure à toute autre, soit (et cela se cumule souvent) parce que les chercheurs s'intéressent aux données linguistiques en « oubliant » que celles-ci ne sont pas données mais produites dans un contexte social par des personnes elles aussi socialement situées.

- projets pour lesquels le recueil de consentement se réduit à une simple « demande d'autorisation » quand celle-ci n'est pas considérée comme purement implicite.

- projets engagés dans de « bonnes pratiques » et utilisant un formulaire standard validé par une autorité juridique.

- projets dont les responsables souhaitent se protéger contre d'éventuelles plaintes ou recours de participants (notamment dans le cadre de projets fondés sur des expériences en laboratoire).

Très rares sont les projets qui développent une véritable réflexion sur les aspects éthiques mais aussi sur les effets méthodologiques voire théoriques produits par le recueil de consentement. On les trouve principalement en ethnolinguistique et en analyse de la conversation.

Prenant en compte la diversité des pratiques, nous n'avons pas souhaité fournir des formulaires clefs en main qui dispenseraient les chercheurs d'une réflexion et d'une description de leurs pratiques et objectifs. Le *Guide des bonnes pratiques* détaille les éléments permettant cette réflexion mais aussi l'appropriation par les chercheurs des enjeux et contenus juridiques afin de faciliter le recueil de consentement éclairé.

DEFINITION DU « CONSENTEMENT ECLAIRE »

On parle souvent de formulaires d'autorisation à soumettre aux informateurs ; il est cependant important de faire dépendre cette autorisation de l'information préalable donnée aux personnes concernées : sans *information*, la *demande d'autorisation* n'a pas d'objet ni de sens. C'est pourquoi on parle de *consentement éclairé (informed consent*), dans le sens où l'acceptation de l'enregistrement est étroitement dépendante de la compréhension des finalités pour lesquelles il est effectué. Sur certains terrains, la difficulté de faire comprendre les finalités de la recherche ne doit cependant pas inciter le chercheur à passer outre la demande de consentement, et celle-ci doit alors être formulée en accord avec le type de société dans laquelle se déroule le terrain (par exemple, comment concevoir un consentement individuel signé dans une société à tradition orale dans laquelle le droit privé n'a aucun sens ?). (GBP :60)

QU'EST-CE QU'*INFORMER* ?

Au cœur du consentement éclairé, il y a l'exigence d'informer les participants enregistrés. Toutefois, dès que l'on interroge cette exigence, les questions surgissent. Qu'est-ce qu'« informer » ? Informer « à propos de quoi » ? A quelles conditions peut-on dire que cette information produit le statut « éclairé » de son destinataire ?

La notion même d'« information » peut laisser penser à un simple transfert de messages et de contenus ; elle tend à gommer les processus, les contextes et les contingences qui caractérisent cette activité communicationnelle par laquelle un enquêteur explique l'objet de son enquête à ses partenaires sur le terrain. Dès que l'on réfléchit en termes de type d'activité, l'« information » aux enquêtés pose une série de problèmes à résoudre :

– l'adéquation au destinataire : l'explication du projet de recherche, pour être comprise et partagée, demande à être ajustée aux compétences, au niveau de langue et de compréhension du destinataire, cet ajustement concerne aussi le contexte et les modalités

de l'enquête, prenant en compte l'adéquation entre ce que les partenaires voient faire sur le terrain et les explications qu'on en donne ;

– l'explicitation des finalités de l'enquête doit se faire sans nuire à celle-ci : cela pose la question de l'équilibre à trouver entre la transparence de l'enquête et les transformations qu'elle peut induire sur les conduites des participants ;

– l'explication du projet de recherche peut se faire à des niveaux de généralité différents (de « c'est une enquête sur les façons de parler des gens » à « c'est une enquête sur la fréquence et les contextes de la liaison non obligatoire en français »).

L'information aux enquêtés comprend non seulement des explications du projet scientifique mais aussi des informations précises concernant par exemple :

– les responsables de l'enquête et leur affiliation institutionnelle, ainsi que les financeurs,

– une adresse de contact,

– les personnes qui auront accès aux données et qui travailleront sur elles,

– la façon dont les enquêtés ont été choisis et la population dont ils font partie,

– la façon dont les données seront anonymisées,

– le fait que les données seront transcrites selon des conventions particulières (possibilité de donner un exemple),

– la façon dont les données seront archivées une fois l'enquête terminée (conservation ou destruction à la fin de l'enquête, conservation auprès de quel garant, modalités de réutilisation éventuelle, transmission à d'autres chercheurs),

– les modalités d'accès aux informations relatives au projet et concernant tout particulièrement les données/analyses faisant référence à la personne (possibilité d'accès aux fichiers et informations concernant tout particulièrement la personne),

– les droits de la personne, notamment le droit de rétractation,

– les risques éventuels ainsi que les retombées positives, morales ou matérielles, de l'étude.

Les modalités d'information peuvent, elles aussi, varier selon la culture des destinataires, en particulier :

– l'information peut se faire de manière orale : individuellement dans des conversations familières, collectivement dans des réunions d'information…

– elle peut se faire de manière écrite (par une brochure, un dépliant…) ou par courriel.

Dans le contexte d'une culture écrite, il est recommandé de laisser un texte ; de même, l'indication d'un site Internet où suivre l'évolution du projet (éventuellement avec des modes d'accès particuliers) peut être utile.

4.3.7. Anonymiser [ex :ESLO_anonymisation] [retour]

L'anonymisation des données est surement la pratique, en réponse à des problèmes juridiques, la plus courante dans les recherches en SHS mais c'est aussi celle qui est la moins

maitrisée par cette communauté scientifique. C'est un bel exemple de pratiques scientifiques fondées sur un savoir populaire qui n'est pas interrogée lors de la démarche de collecte et d'exploitation des données.

Il faut d'abord préciser que dans un cadre juridique, l'anonymisation n'est pas obligatoire. Elle l'est seulement dans le cas où il y a des données identifiantes (nous reviendrons sur cette notion) utilisées sans l'autorisation des personnes concernées. Il s'agit surtout d'une alternative au défaut de consentement :

> « *L'anonymisation des données est une garantie importante en matière de légalité des données et de leur usage ; dans certains cas, si elle garantit véritablement la non-identification des personnes concernées, et si par ailleurs les données ne sont pas protégées par le droit d'auteur, elle peut permettre d'utiliser des données même en l'absence de demande d'autorisation préalable.* » (GBP:67)

Sur un plan éthique, l'anonymat est au cœur d'une polémique entre les défenseurs de l'anonymat et ceux qui au contraire plaident pour une « dénomination » des locuteurs considérés comme des partenaires du projet (https://ethiquedroit.hypotheses.org/). Là encore nous pouvons percevoir l'impact d'une considération méthodologique ou éthique, en l'occurrence sur la catégorisation des éléments d'un corpus.

L'anonymisation n'est donc pas une opération systématique mais doit résulter d'une approche réflexive en termes juridiques mais aussi théoriques et méthodologiques. Avant de développer ces aspects il convient de préciser la notion même d'anonymat :

> *« Bien qu'on parle souvent d'anonymisation, la question légale qui se pose est celle de l'impossibilité d'identifier des personnes : l'enjeu est que, sur la base des données recueillies et de leurs modes de représentation (transcription par exemple), on ne puisse pas identifier les personnes concernées. »* (GBP:67)

Pour les linguistes, l'anonymisation a consisté pendant longtemps à une simple opération de masquage ou de codification des noms propres. Le cadre juridique dépasse largement cette opération en pointant la « non identification » des personnes. D'un point de vue linguistique cette approche permet de formuler un double constat : contrairement à ce que nous pensions, les noms propres ne correspondent pas complètement à des éléments construisant une relation avec un référent unique (il existe de nombreux Jacques Martin) et deuxièmement la langue contient de nombreux éléments permettant de construire cette relation, éléments qui fonctionnent le plus souvent en faisceaux contextuels (Eshkol *et al.* 2015).

Ainsi il convient de parler de procédure d'identification :

> *« Les procédures d'identification sont bouleversées par les technologies actuelles qui offrent des facilités de stockage et de diffusion des données, mais aussi de puissants outils de traitement des informations (tri, recoupement, requêtes croisées…).* (GBP :67)

Ainsi non seulement la liste des éléments qui permettent une identification est plus importante que la simple action de « non nomination » au sens propre, mais la question des moyens utilisables pour cette identification est prépondérante. C'est d'ailleurs ce qui est précisé dans la directive 95/46/CE[133] :

« *…pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne* ».

Sur ce point, la situation a déjà fortement évoluée entre 2006 et 2015. Les avancées technologiques du traitement des « big data » sont un enjeu tel qu'il est extrêmement difficile d'imaginer quels seront les moyens disponibles dans les années à venir.

L'anonymisation est relativisée par différents facteurs intervenant soit lors de la production des données – et selon les spécificités de ce qui advient durant l'enregistrement –, soit lors de la réception de ces données :

- L'anonymisation opère d'abord sur une série de formes censées contenir les indications principales permettant l'identification de la personne ; néanmoins n'importe quelle référence ou forme peut, selon les contextes, conduire à l'identification de la personne, et souvent d'une manière qui passe au premier abord inaperçue pour l'enquêteur. (…)
- Le caractère reconnaissable de ces détails dépend de manière cruciale du contexte de réception et plus spécifiquement du public qui consultera ou prendra connaissance des corpus. (…) La valeur identifiante d'un détail dépend donc du contexte de réception des données.
- D'autres aspects sont liés au *recoupement* d'informations venant de plusieurs sources (cela peut concerner par exemple la relation entre données anonymisées et métadonnées).

Les possibilités technologiques sont couplées avec la maturité de la recherche en matière d'éthique. Pour s'en convaincre, il suffit de lire les quelques lignes consacrées à ce sujet dans le livre qui détaille la méthodologie suivie lors de la constitution du corpus du Français fondamental :

> *« Dans la transcription nous avons eu soin de remplacer par des initiales les noms des personnes mises en cause* » (Gougenheim *et al.*, 1956 :64[134]).

L'identification automatique des visages à des fins de reconnaissance des personnes est un exemple de ces avancées qui doit interpeler les chercheurs responsables de la collecte du

[133] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:fr:HTML

[134] GOUGENHEIM, G. (1956). *L'élaboration du français élémentaire: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*.

consentement et de l'anonymisation. En effet, l'anonymat à moyen terme semble être une utopie difficilement maitrisable dans le Web 3.0.

Ces considérations concernent :

> *« - tout ce qui permet d'identifier directement une personne : par référence au locuteur ou à un tiers et à sa sphère privée, sur la base des manifestations du locuteur, comme sa voix ou son apparence physique ;*
> *- tout ce qui peut lui porter préjudice ;*
> *- tout ce qui peut indirectement permettre, par recoupement d'informations, de remonter au locuteur concerné.*
>
> *Les opérations qui suppriment ces références ou ces manifestations sont appelées des procédures d'« anonymisation » des données. » (GBP :67)*

Le second point spécifique à l'anonymisation des corpus oraux concerne la diversité des types de données constitutives de l'objet. Ainsi l'anonymisation peut concerner :

- les données primaires vidéo,
- les données primaires audio,
- les données primaires textuelles : documents, officiels ou non, recueillis sur le terrain,
- les données secondaires : transcription, notes de terrain, métadonnées, analyses, descriptions ethnographiques,
- les données secondaires visuelles : copies d'écran (*screen shots*), voire représentations de la voix (oscillogrammes, spectrogrammes…).

Il est difficile – voire impossible – de constituer une liste finie des formes concernées par l'anonymisation. On peut toutefois souligner les formes principales :

> *– « formes nominatives (nom, prénom, surnom ou petit nom, sigle d'entreprise…),*
> *– données personnelles (adresse, numéro de téléphone, numéro de passeport, numéro de compte, âge, lieu de naissance…),*
> *– profession, statut, titres,*
> *– activités sociales,*
> *– parenté, réseaux,*
> *– référence à des lieux (toponymes, institutions, services…),*
> *– référence à des caractéristiques de la personne (physiques, culturelles, médicales…) uniques ou rares dans son milieu,*
> *– caractéristiques physiques : voix, visage, caractéristiques corporelles,*
> *– etc.*
>
> *« L'« etc. » clôturant cette liste souligne le fait que tout élément, selon les contextes d'enregistrement et de réception de cet enregistrement, peut devenir un porteur d'informations sur l'identité des personnes. L'identification des formes concernées par l'anonymisation suppose donc une compétence sociologique et culturelle qui rende le chercheur capable*

> *d'imaginer les usages, les connaissances et les associations qui pourraient permettre l'identification d'une personne sur la base d'une forme donnée. »* (GBP :69)

La dernière phrase de cet extrait du *Guide* représente le travail réalisé avec I. de Lamberterie afin d'ouvrir cette perspective au groupe de travail et donc à l'ouvrage. L'objectif est bien d'accompagner le chercheur vers une approche réflexive et une explicitation de sa démarche afin de lui permettre de construire ses « bonnes pratiques ». Sur la phase d'anonymisation comme sur celle du recueil de consentement, la demande initiale des chercheurs était préférentiellement de trouver des solutions prêtes à l'emploi sous la forme de formulaires à recopier ou de procédure à suivre à la lettre. C'est dans une tout autre approche que des solutions peuvent apparaître au sein d'un projet scientifique. Il y a là encore une relation aux données et à la place de l'observation dans le travail scientifique qui est significative : une conception des données qu'il convient de rendre exploitables, y compris d'un point de vue juridique, avant de les exploiter diffère d'une conception qui refuse de séparer données et analyses. C'est en ce sens que *le Guide* apporte des précisions sur l'ensemble de la méthodologie.

La phase de l'anonymisation est également impliquée dans cette conception des données :

> « *Selon les finalités de l'étude et les contextes de l'enquête, on peut considérer que l'anonymisation doit se faire le plus tôt ou le plus tard possible. La première solution augmente les garanties de confidentialité pour la personne, la seconde maximise les possibilités d'analyse pour le chercheur. Les temporalités peuvent varier selon les types de données aussi :*
>
> *– on évite l'anonymisation sur les données primaires originales de référence car elle pourrait endommager les données elles-mêmes ; par contre les données ainsi non anonymisées doivent être conservées dans un lieu sûr,*
>
> *– les données peuvent/doivent/ne doivent pas (selon les politiques adoptées) être anonymisées lors de leur dépôt pour conservation ; le rôle de garant des institutions assurant la conservation est ici concerné,*
>
> *– on peut travailler (dans un groupe de recherche bien délimité et qui garantit la non circulation externe des données) sur des données non anonymisées et garantir en revanche une anonymisation de tout extrait figurant dans un écrit ou une présentation orale,*
>
> *– on effectue toujours l'anonymisation sur les copies destinées à circuler entre chercheurs extérieurs au projet et parfois entre chercheurs internes au projet (c'est le cas notamment pour de grands consortiums de recherche ou des projets articulant des réseaux d'équipes importants)* ». (GBP:68)

Comment réaliser l'anonymisation ? La question peut être conçue à différents niveaux. Le premier concerne le repérage des éléments à anonymiser. Le projet ESLO a donné lieu à une

tentative d'automatisation de ce repérage (Eshkol *et al*., 2006, 2008, 2011, 2015) dont le bilan est décevant sur un plan technique mais stimulant d'un point de vue scientifique[135].
Le deuxième niveau concerne les formes de remplacement. La forme d'anonymisation généralement adoptée procède par remplacement d'éléments confidentiels par des formes neutres. Ces formes varient selon les supports techniques concernés mais elles ne sont pas sans effet sur la catégorisation même des données comme le démontre le *Guide* sur ces deux exemples :

> – *Remplacement par un hyperonyme ou une abréviation, tel que NN ou NVILLE ou NHOPITAL pour nom, nom de ville, nom d'hôpital, etc. Cette solution peut rester informative (on précise le type de référence de la forme anonymisée). Elle est utile dans les cas où la substitution par pseudonyme (cf. infra ici-même) est impossible, difficile ou non vraisemblable. Cette solution implique le développement de conventions spécifiques pour la notation de ces hyperonymes, qui ne sont pas de même nature que le texte qu'ils remplacent (c'est pourquoi l'emploi des majuscules est parfois proposé, quand il n'entre pas en contradiction avec d'autres emplois de majuscules prévus dans les conventions de transcription).*
> – *Remplacement par un pseudonyme : c'est la solution la plus souvent utilisée, du moins pour les noms de personnes car elle permet une bonne intégration de la forme de remplacement dans le fil du discours, n'attire pas l'attention sur elle, est vraisemblable et garde un certain nombre d'indications contenues dans la forme initiale. Cela n'est toutefois possible que si le choix des pseudonymes est réfléchi et répond aux problèmes suivants : le pseudonyme est choisi dans le même champ paradigmatique que la forme qu'il remplace (par exemple « Ahmed » sera remplacé par « Moustapha » plutôt que par « Albert », le pseudonyme tentant de conserver des traits d'ethnicité), dans certains cas, notamment si l'interaction enregistrée le rend pertinent, on veillera à conserver les connotations possibles du nom (par ex. s'il est à la base de plaisanteries ou de jeux de mots) et le nombre de syllabes et certaines caractéristiques phonétiques et prosodiques (si elles sont exploitées dans l'interaction) (…)* (GBP :70)

Enfin un troisième niveau concerne les possibilités technologiques de remplacement des formes à anonymiser. L'exemple d'utilisation par le projet ESLO du script développé par D. Hirst permet de mesurer les fonctionnalités que peuvent offrir des outils qui associent la transcription synchronisée, l'annotation et le traitement du signal.

La phase d'anonymisation est donc également une phase de transformation du terrain en un espace domestiqué pour les besoins de la recherche en adéquation avec un cadre juridique.

[135] Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO https://halshs.archives-ouvertes.fr/halshs-01165957

4.3.8 Bref bilan [retour]

Le *Guide* a été tiré à 2000 exemplaires et largement diffusé. Il est également disponible en trois langues sous une forme électronique :

Version française : https://hal.archives-ouvertes.fr/hal-00357706
Version anglaise : https://halshs.archives-ouvertes.fr/halshs-01165893
Version allemande : https://halshs.archives-ouvertes.fr/halshs-01165896

Il a également fait l'objet d'une traduction en coréen.

La réception a été fructueuse au sein de la communauté des linguistes. La critique principale entendue à l'occasion de différentes présentations qui ont été faites de l'ouvrage a trait à l'absence de réponses sous formes de formulaires ou de procédures standardisés. Pourtant le *Guide* s'inscrit dans une double perspective explicite : (i) il n'existe pas de réponse juridique simple face à un objet complexe et des pratiques hétérogènes et (ii) seule l'approche réflexive peut permettre au chercheur d'appréhender les réponses adéquates à construire pour chacun de ses projets.

Une réflexion pour quiconque penserait que l'activité scientifique est indépendante des effets de champ. Depuis la parution du *Guide*, différents projets ont pris pour objet les aspects juridiques et éthiques en linguistique et plus largement en SHS. Dans chaque cas, il semble que les responsables aient préféré ignorer ce qui avait été fait. Ce n'est ni une posture d'opposition, ni une attitude critique à l'encontre du *Guide,* ce qui présenterait l'avantage de faire avancer l'état de la science sur le sujet*,* mais plutôt la volonté de prendre possession d'un champ nouveau sans assumer les exigences scientifiques d'une démarche critique et cumulative.

4.4 Corpus de la parole [retour]

4.4.1 Origine et buts [cadre théorique et politique]

Le programme Corpus de la parole est né parallèlement au groupe de travail sur le *Guide des bonnes pratiques*.
Son origine est directement liée au pilotage du conseil scientifique de l'Observatoire des pratiques linguistiques de la DGLFLF ainsi qu'au cadre théorique de la sociolinguistique tel que je le défends au sein du Ministère de la Culture et de la Communication.
Plus précisément, six facteurs ont déclenché ce programme qui a pris une ampleur croissante au sein de la DGLFLF :

1. La recherche d'un programme fédérateur qui permette un partenariat entre le MCC et le Ministère de l'Enseignement Supérieur et de la Recherche, et plus précisément entre la DGLFLF, la BnF, les Fédérations de recherche en linguistique et les nouvelles structures dédiées à l'archivage scientifique et aux ressources numériques.
2. La volonté d'inscrire les langues dans les actions concrètes du Ministère de la Culture, notamment dans le cadre du Plan de numérisation.
3. Le souhait de répondre à des besoins exprimés par les équipes de recherche en linguistique qui travaillent sur les langues parlées sur le territoire français.
4. La volonté de la DGLFLF de développer des initiatives en faveur des « langues de France » puisque cette nouvelle notion se trouvait introduite dans le champ de la politique linguistique à la suite de la signature par la France de la Charte de protection des langues régionales et minoritaires.
5. La volonté de reconnaitre les pratiques linguistiques orales à la fois comme un objet scientifique légitime et comme un objet patrimonial tout aussi légitime.
6. L'opportunité de mettre en pratique les recommandations du *Guide des bonnes pratiques* et d'accompagner la linguistique de l'oral dans le « tournant » des humanités numériques.

Les objectifs du programme sont présentés dans la note que j'ai rédigée en janvier 2006 (Baude, note interne DGLFLF, 2006[136]) et qui a été reprise dans différents documents et publications[137].

[136] https://www.nakala.fr/nakala/data/11280/a55a0199
[137] La reprise quasi systématiques des mêmes phrases dans les différents documents est due au statut de document de référence, validé après un long circuit interne à la DGLFLF, du conseil scientifique au Délégué général et au Cabinet du Ministre.

Presque dix ans après le début de ce programme, nous retrouvons dans la note de présentation les points principaux qui structurent celui-ci :

Premièrement la reconnaissance de la diversité linguistique et de la richesse de la variation linguistique constatées dans les pratiques même des locuteurs. Cette reconnaissance s'appuie principalement sur la notion de « langue de France » telle que définie par P. Encrevé et B. Cerquiglini dans la position française sur la Charte européenne :

> « *La France dispose d'une richesse linguistique fondée sur la diversité. A côté du français, langue nationale, présent sur les cinq continents, les langues de France constituent un patrimoine culturel unique. Une grande diversité les caractérise : langues romanes, langues germaniques, breton (celtique), basque (non indo-européen) dans l'Hexagone ; créoles, langues amérindiennes, polynésiennes, austronésiennes, bantoue outre-mer ; elles sont parlées par un nombre très variable de citoyens : si l'arabe compte 3 ou 4 millions de locuteurs en France, le neku ou l'arhâ n'en comptent que quelques dizaines, en Nouvelle-Calédonie. Entre les deux, les différents créoles, ou le berbère, sont parlés par près de 2 millions de Français* ». (Baude, Note2066 CP:1)

Particularité qui relève du cadre théorique variationniste validé par le conseil scientifique, le « français parlé », comme nous le verrons dans le paragraphe suivant, nécessitera les mêmes dispositions de reconnaissance que les langues de France. Sur ce point, dialectologie et sociolinguistique se rejoignent pour peu qu'on se place du côté de l'observation d'un marché linguistique.

Deuxièmement, cette reconnaissance est présentée à partir des archives sonores, de leur conservation et de leur diffusion. Depuis le commencement, cette reconnaissance ne se situe pas dans le cadre d'une patrimonialisation d'archives de pratiques disparues qu'il conviendrait de préserver dans un esprit muséologique, mais plutôt dans une reconnaissance de la collecte de pratiques linguistiques comme ressources vivantes et qui ont une fonction à remplir dans des actions de politique sociale, éducative et culturelle :

> « *Ce patrimoine est méconnu, et si des archives sonores existent pour la quasi-totalité de ces langues, force est de reconnaitre que cette richesse, constituée du français parlé et de la diversité des langues de France, n'est accessible ni à l'ensemble de la communauté scientifique, ni au grand public. Plus grave encore, de nombreux documents sonores uniques, conservés sur des supports physiques en fin de vie (bandes magnétiques), sont voués à disparaitre à tout jamais dans un délai très bref. Il s'agit souvent des derniers et seuls documents sur des langues de France (langues de Guyane, de Nouvelle-Calédonie…), et même sur le français (la DGLFLF a*

> *numérisé les seuls enregistrements de français constitués par des linguistes dans les années 70). La numérisation offre non seulement la possibilité de sauver ces documents, mais aussi de les valoriser en les transformant en de véritables ressources linguistiques numériques, assurant ainsi la vitalité de cette diversité*. » (Baude, Note2066 CP:1)

Troisièmement, ces archives sont définies comme des « corpus » que le passage au numérique a radicalement transformées en tant qu'objet scientifique et patrimonial :

> *« Un corpus oral n'est pas en effet une simple collection d'enregistrements de la parole humaine, mais un objet "construit", (enregistrements + catalogage, indexation, transcription, synchronisation du son et de la transcription...) : ce sont la numérisation, la transcription, l'élaboration de métadonnées... qui permettent de passer d'un simple enregistrement à un objet patrimonial pouvant faire l'objet de recherche et de valorisation (par exemple, s'agissant de la parole, il est techniquement impossible de faire de la recherche d'occurrences sur du son, ce n'est possible que sur une transcription). »* (Baude, Note2066 CP:1)

Quatrièmement, le développement de corpus oraux est présenté comme une initiative de politique scientifique et culturelle majeure. Nous retrouvons les traces d'une lutte interne à la DGLFLF sur les orientations d'une politique linguistique qui oscille entre la défense du français et la reconnaissance de la diversité linguistique :

> « *Ainsi, le développement des corpus oraux (collections ordonnées d'enregistrements de productions linguistiques orales, et multi-modales du type LSF) de français et des langues parlées en France est actuellement un enjeu capital pour la politique linguistique de la France. Alors que la plupart des langues européennes disposent de corpus oraux accessibles en ligne, et souvent gratuitement, un tel outil n'existe pas dans notre pays, ce qui a des conséquences néfastes pour la visibilité et la vitalité du français et des langues de France. C'est un enjeu pour la recherche linguistique et pour le développement de l'ingénierie linguistique (reconnaissance et synthèse de la parole, traitement automatique des langues), c'est un enjeu aussi pour l'enseignement de ces langues, pour la sauvegarde et la diffusion du patrimoine oral. »* (Baude, Note2066 CP:1)

Enfin cet enjeu est présenté comme dépassant les frontières de la métropole et aussi de la nation et de l'Europe. L'objectif est de pouvoir mobiliser des moyens financiers – mais pas seulement – conséquents :

> *« En tout état de cause, un développement de ce programme à la hauteur des besoins et de l'enjeu nécessiterait des sommes nettement plus importantes. En effet, ce programme doit permettre de pallier le retard de la France dans la diffusion du patrimoine linguistique numérique national, mais surtout de proposer une initiative, unique en Europe et au niveau international, de numérisation du patrimoine dans le respect des nouvelles technologies de conservation, mais aussi d'enseignement et de traitement automatique des langues, et d'assurer ainsi la vitalité du français et des langues de France, véritable source de diversité culturelle. »* (Baude, Note2066 CP:2)

En tirant parti de ces arguments, trois actions ont été proposées pour mener à bien la première phase de ce programme : l'élaboration d'un partenariat MESR-MCC, un travail concret sur les « bonnes pratiques » et des actions de numérisation des corpus oraux.

Le partenariat avec le MESR a concerné dans un premier temps un soutien aux actions de recherche sur ou à partir de corpus oraux en liaison avec les fédérations de recherche en linguistique du CNRS qui regroupent la totalité des UMR et quelques équipes universitaires travaillant en linguistique. L'objectif est avant tout de construire un réseau autour des corpus oraux dans un paysage institutionnel où la linguistique de corpus est un domaine encore très éloigné des travaux sur les pratiques linguistiques :

> « 1° *Un soutien à la recherche (constitution et exploitation de ressources linguistiques sonores), partenariat avec les fédérations des laboratoires de recherche en linguistique du CNRS (Institut de linguistique française, ILF-FR 2393, et Typologie et Universaux Linguistiques, TUL-FR 2559) pour la sauvegarde et le développement des corpus oraux. Ce partenariat s'est traduit par une aide globale de 69 000 € à ces deux fédérations en 2004, et de 40 000 en 2005. Ces sommes, à la mesure des moyens de la DGLFLF, sont modestes, pour ne pas dire symboliques, au regard des besoins ; mais cette action a permis de motiver très fortement les différents acteurs et d'orienter la recherche vers un objectif de mise à disposition de données représentant la diversité du patrimoine linguistique. »* (Baude, Note2066 CP:2)

La seconde action s'appuie sur le travail commencé autour des aspects juridiques qui semble un domaine émergent et prometteur à cette époque.

> « 2° *Le Guide des bonnes pratiques. La création d'un groupe de travail comprenant des linguistes (CNRS et Université), des juristes et des conservateurs (BnF, INA, Archives), pour réfléchir sur les questions théoriques et méthodologiques relatives à la numérisation et à*

> *l'exploitation des corpus oraux, a abouti à la rédaction d'un "Guide des bonnes pratiques", à la fois juridique, éthique et technique, à paraitre en mars 2006 aux éditions du CNRS. Ce Guide a donné lieu à une journée d'étude en mai 2005 à la BnF et sa version provisoire, fort bien accueillie par les chercheurs et les conservateurs, fait déjà office de référence en la matière*. » (Baude, Note2066 CP:2)

Enfin la troisième action immédiate concerne la numérisation de corpus oraux. Il s'agit ici d'une opportunité car les plans de numérisation bénéficient encore d'un large soutien et l'absence de travaux antérieurs sur les langues facilite la prise en compte d'un programme linguistique.

> « 3° *La numérisation d'archives linguistiques sonores. Dans le cadre du plan de numérisation piloté par la MRT (Mission pour la Recherche et la Technologie) du ministère, la DGLFLF a présenté un programme consistant à numériser des fonds sonores du français et des langues parlées en France (numérisation des fonds fragiles dont les supports analogiques sont dans un état de détérioration ; numérisation de fonds plus récents pour permettre leur intégration dans une base de données ; indexation, catalogage et établissement de normes d'interopérabilité), à les valoriser par la création d'un site portail présentant les corpus de français et de langues de France, et à intégrer dans ce site une base de données regroupant une riche collection de corpus desdites langues. Cette base de données permettra une mise à disposition de ressources représentant la diversité des pratiques linguistiques en France. La demande initiale était de 200 000 € pour l'année 2005 ; le projet a été retenu à hauteur de 85 000 €, ce qui permet de lancer la première tranche du projet. »* (Baude, Note2066 CP:2)

Cette troisième action va nécessiter la création d'un entrepôt de corpus afin de permettre leur archivage et leur accessibilité. Ce sera un projet délicat puisqu'à l'époque, il n'existe pas encore d'infrastructures dédiées aux données scientifiques numériques.

La réponse institutionnelle du CNRS, sollicité par le MCC pour mener à bien les opérations du plan de numérisation, sera de renvoyer vers l'un des futurs centres de ressources numériques pour l'oral : le CRDO-Paris. C'est celui-ci qui, naissant, accompagnera les balbutiements du programme Corpus de la parole dans l'élaboration d'un entrepôt de corpus oraux.

Par la suite, le TGE ADONIS confirmera la mission confiée au CRDO et développera un projet pilote (de 2008 à 2010) sur l'archivage pérenne des données orales en regroupant les centres de calcul du CINES et de l'IN2P3, le CRDO, la DAF. Ce projet donnera lieu à *un Guide*

méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles.

Le programme Corpus de la parole se situe en amont de toutes ces initiatives. Il figure à ce titre un projet expérimental d'envergure qui bénéficiera de la création de ces infrastructures mais sera aussi confronté à des difficultés dues à la précocité du projet.

Ainsi l'objectif du programme était double :

- Numériser, afin de les préserver et les rendre accessibles, des corpus sonores enregistrés sur des supports analogiques en fin de vie.

- Accompagner un changement des pratiques des chercheurs qui collectent et analysent des corpus oraux à l'ère de ce qui deviendra les « humanités numériques ».

4.4.2 Inventorier et mutualiser des corpus [retour]

Au début des années 2000, la situation des corpus oraux en France n'était pas connue. L'expérience du Corpus d'Orléans (ESLO) et d'autres semblables laissait présager la déréliction des enregistrements stockés sur des bandes magnétiques et autres supports analogiques. J'ai donc proposé au Conseil scientifique de la DGLFLF de commander un inventaire des corpus oraux en France. Celui-ci fut réalisé par Paul Capeau et Magali Seijido en 2005. Notons qu'un second inventaire des corpus francophones, avec leur analyse détaillée, a été réalisé par F. Gadet en 2013. Un inventaire ouvert de 2011 à 2015 par le consortium IRCOM de la TGIR Huma-Num complète cette liste.

Ces inventaires démontrent la diversité des corpus existants mais aussi l'hétérogénéité des données. Le plan de numérisation fut donc l'occasion de définir des formats, normes et pratiques afin de préserver et diffuser cette ressource.

La première étape fut celle de l'élaboration de préconisations pour la numérisation des enregistrements, pour les transcriptions et pour la description des ressources

Préconisations pour la numérisation des enregistrements :

> *« Pour la numérisation des anciens supports analogiques, le CRDO-Paris a défini des critères de qualité minimaux. Ces critères, inspirés de ceux préconisés par IASA[138] (IASA 2009) ont été validés par le conseil scientifique du programme Corpus de la parole en accord avec le département des archives sonores de la BnF et communiqués, via les fédérations de linguistique, aux chercheurs et laboratoires qui pratiquaient eux-mêmes la numérisation. Il s'agissait d'une préconisation contractuelle qui a donné lieu à l'élaboration d'une annexe technique, systématiquement présente dans les conventions de la DGLFLF. Le CRDO-Paris, qui pilotait également une*

[138] *International Association of Sound and Audiovisual Archives*.

partie des numérisations pour le compte des chercheurs et laboratoires qui le souhaitaient, appliquait aussi obligatoirement ces préconisations lors des opérations de numérisation à l'aide des équipements d'un laboratoire qui s'en était doté pour ses besoins propres (le LACITO[139]). Pour les enregistrements audio ces préconisations étaient les suivantes : échantillonnage 44,1 kHz au minimum (96 kHz au LACITO) ; quantification : 16 bits au minimum (24 bits au LACITO) ; copie droite sans retouche ; format WAV ; encodage : PCM[140]. » (Baude et Jacobson 2011 :52)

Préconisations pour l'écriture des transcriptions :

Pour les annotations pouvant accompagner les enregistrements141, le CRDO-Paris a défini, toujours après validation du conseil scientifique, des recommandations allant jusqu'à un modèle cible en XML. Ce modèle, exprimé dans une DTD XML, définit une structure minimale permettant :

de coder la transcription d'un enregistrement ;

d'ajouter une traduction en français (ce qui été demandé pour les langues autres que le français ou pour des transcriptions non orthographiques du français) ;

de découper la transcription en segments (phrase ou groupe de souffle) ;

de noter les repères temporels de début et de fin des segments.

Quelques raffinements du modèle permettent également d'indiquer le locuteur (utile pour les dialogues), le type de transcription (orthographique, phonétique, phonologique).

Ce modèle minimal peut être atteint soit directement, soit en passant par des formats et outils qui permettent de faire une annotation plus riche et plus fine. En particulier, les formats en sortie des outils Transcriber142 et ITE143 peuvent être directement exploités dans le cadre du projet, la transformation vers le format cible étant alors complètement automatisée. D'autres formats, tels que ceux utilisés par les outils CLAN ou ELAN, doivent faire l'objet d'une normalisation afin d'être transformés de manière souvent ad hoc dans le format cible. Dans ce dernier cas, les deux formats sont conservés : le format d'origine et le format cible.

Ces recommandations ne portent que sur la forme à utiliser pour exprimer les transcriptions. Aucune indication ni directive n'est donnée pour expliquer aux chercheurs comment ils doivent transcrire leurs enregistrements et en donner une traduction en français. Les linguistes ont parfois établi des conventions accompagnées de manuels144, mais d'une langue à l'autre (du français à la langue des signes française, par exemple), ou d'un domaine linguistique à un autre (de la phonétique à la dialectologie, par exemple), ces conventions sont peu partagées. Il est en revanche conseillé d'identifier, sous forme d'une référence dans les

[139] Laboratoire de langues et civilisations à tradition orale.
[140] Pulse-code modulation. Il s'agit d'un codage sans compression.
[141] Le programme de la DGLFLF prévoit également une phase de valorisation des enregistrements à l'aide d'une ou plusieurs couches d'annotations (transcription, glose, traduction, annotations morphosyntaxiques, syntaxiques ou autres).
[142] Transcriber (http://trans.sourceforge.net/).
[143] ITE Interlinear Text Editor (http://michel.jacobson.free.fr/ITE/).
[144] Conventions pour ESLO (http://eslo.in2p3.fr/).

> *métadonnées, la ou les ressources qui décrivent de manière explicite l'ensemble des conventions utilisées.* (Baude et Jacobson 2011 :53)

Préconisation pour la description des ressources :

> *La description des ressources (les enregistrements et les transcriptions) a également donné lieu à l'élaboration de préconisations. Cette description repose sur un jeu de métadonnées qui doivent suivre le schéma XML défini par OLAC145. Ce schéma reprend ceux du Dublin-core et du Dublin-core qualifié146 auxquels sont ajoutés cinq attributs associés à des vocabulaires contrôlés (role, language, linguistic-field, linguistic-type et discourse-type). Les recommandations précisent la manière d'utiliser ce schéma OLAC pour décrire les ressources (les éléments obligatoires, facultatifs, des explications, des exemples, etc.).* (Baude et Jacobson 2011 :54)

Ces trois préconisations peuvent sembler élémentaires. Une découverte surprenante, et une grande difficulté du programme, fut la difficulté éprouvée par certains producteurs de corpus pour se conformer à celles-ci. Avec plusieurs années d'expérience, nous pouvons esquisser une brève typologie des comportements des producteurs de corpus oraux impliqués dans ce programme :

- les chercheurs qui maitrisent déjà ces considérations techniques et méthodologiques et qui structurent, décrivent, encodent et formatent leur corpus de cette manière ou d'une manière similaire au sein de leur laboratoire ou d'un consortium plus large (MSH, Centre de ressources…),

- les chercheurs qui ont déjà géré et structuré leurs corpus avec suffisamment de rigueur pour se conformer rapidement à ces préconisations,

- les chercheurs novices en ce domaine qui étaient en attente de préconisations pour s'y conformer,

- les chercheurs dont les corpus n'ont jamais, pour diverses raisons, été prévus pour être conservés et diffusés et pour lesquels l'investissement en temps humain pour atteindre ces objectifs est considérable,

- les chercheurs qui, pour diverses raisons, affichent une volonté de conserver et diffuser leurs corpus mais qui n'iront pas, volontairement, au bout de cette démarche.

[145] *Open Language Archives Comunity*.
[146] Dublin-core ou norme ISO 15836.

4.4.3 L'élaboration d'un « entrepôt de corpus » *[cadre théorique]* *[retour]*

Afin de gérer la conservation et l'accessibilité de ces corpus, le CRDO-Paris (devenu COCOON en 2012) conformément aux travaux du projet pilote sur l'archivage prenne de l'oral, a mis en place une architecture technologique selon le modèle OAI et la norme OAIS :

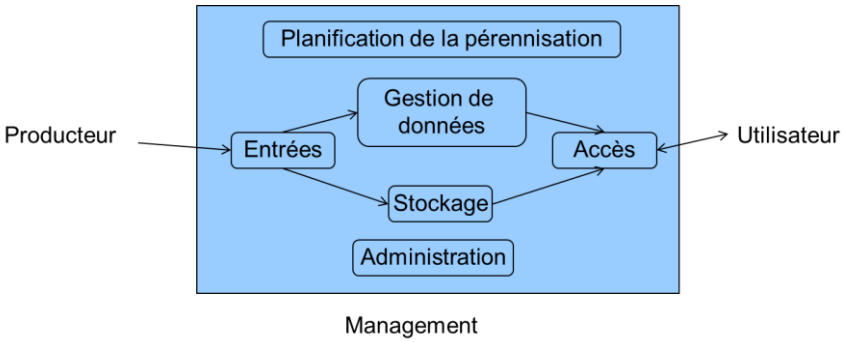
Le mouvement OAI a pris racine lors d'un meeting les 21 et 22 octobre 1999 à Santa Fe (Nouveau Mexique). Ce meeting réunissait les plus grands archivistes de pre-prints et de e-prints (arXiv.org, CogPrints, NDLTD, RePEc, etc.). Son objectif était de faire discuter entre eux ces acteurs sur les difficultés ou impossibilités d'effectuer des requêtes portant sur plusieurs de leurs archives en même temps. Des solutions devaient être proposées pour résoudre ce problème d'interopérabilité.
Pour ce meeting, un prototype d'architecture avait été établi: l'UPS (Universel Preprint Service). Depuis ce prototype a évolué et est connu sous le nom de OAI-PMH (Open Archive Initiative - Protocol for Metadata Harvesting), et l'organisation qui le gère sous le nom de OAI (Open Archive Initiative).
Dans une architecture OAI on distingue deux rôles principaux que sont les fournisseurs de ressources et les fournisseurs de services. COCOON est typiquement un fournisseur de ressources. Ses ressources ont pour seule particularité d'être des ressources d'une communauté particulière et d'une nature particulière (des corpus oraux). Les fournisseurs de services offrent des outils pour un ou plusieurs fournisseurs de ressources. L'outil le plus fréquemment proposé est le moteur de recherche.
Le dialogue entre un fournisseur de service et un fournisseur de ressources se fait à l'aide du protocole défini par l'OAI (OAI-PMH). Les informations qui sont véhiculées lors de ces dialogues sont des métadonnées et non pas les données elles-mêmes, c'est-à-dire des descripteurs de ressources (l'équivalent d'une notice bibliographique pour la publication papier classique)
L'entrepôt OAI COCOON
COCOON est organisé sur le modèle OAI comme un fournisseur de ressources (ou "entrepôt"). L'interface qui répond au protocole OAI-PMH est accessible à partir de l'URL de base http://cocoon.huma-num.fr/crdo_servlet/oai-pmh. C'est cette URL qui est utilisée par les fournisseurs de services pour moissonner les métadonnées de COCOON. Seules les ressources dont les métadonnées sont librement accessibles sont moissonnables par ce protocole. Cette interface permet de délivrer au choix, les métadonnées suivant plusieurs modèles (correspondant à des schémas de métadonnées différents) :

- 'oai_dc': correspond au Dublin-core simple (norme ISO 15836). Ce schéma, obligatoire dans le cadre de l'OAI, permet d'assurer la plus large interopérabilité.
- 'crdo_dcq': correspond au Dublin-core Qualifié.
- 'olac': correspond au schéma défini par l'organisation "Open Language Archive Community", qui est basé sur le Dublin-core qualifié avec quelques ajouts de vocabulaires contrôlés.
- 'mods': "Metadata Object Description Schema" est un schéma de description de données bibliographiques développé à l'origine par la Bibliothèque du Congrès (Washington). Ce schéma permet d'assurer une base de compatibilité avec les outils bibliographiques.

L'OAIS correspond à la norme ISO 14721:2003 (révisé en 2012) - "Reference Model for an Open Archival Information System". Il s'agit du résultat d'un groupe de travail du Consultative Committee for Space Data Systems (CCSDS) rassemblant au départ les grandes agences spatiales et élargi pour des raisons de généricité du sujet à d'autres domaines comme celui des archives institutionnelles ou des bibliothèques.
L'OAIS présente un modèle conceptuel circonscrivant l'organisation d'un système d'archivage. Il définit:

- Un modèle fonctionnel comportant les différentes entités (l'entrée, le stockage, la gestion des données, l'administration, la planification de la préservation et l'accès);
- Des acteurs (les producteurs, l'archive, les utilisateurs, le management);
- Un modèle d'information avec l'information de représentation (syntaxique et sémantique), l'information de pérennisation (identification, provenance, contexte, intégrité) et l'information d'empaquetage;
- Une définition des paquets d'information: les SIP – Submission Information Packages ou Paquets d'informations à verser –, les AIP – Archival Information Packages ou Paquets d'informations archivées –, les DIP – Dissemination Information Packages ou Paquets d'informations diffusées.
- Un lexique associé à ces concepts.

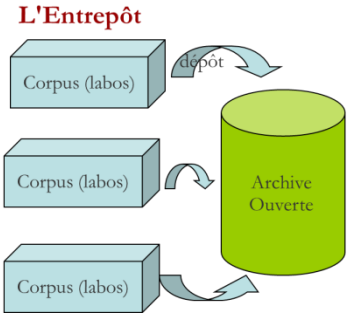
Le schéma suivant présente ce modèle utilisable dans le cas qui nous concerne :



Appliquée au programme Corpus de la parole, l'organisation mise en place est la suivante : Le CRDO (2005-2011) recueille les corpus et les organise dans une "archive ouverte" (au sens OAI). Dans cette archive, les données sont normalisées :

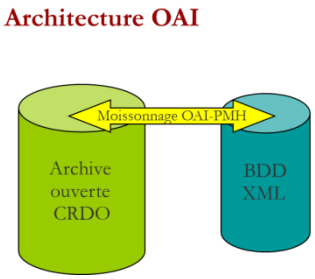
- les annotations en xml,
- les enregistrements audio en wav/pcm,

Les ressources sur le "français et les langues de France" constituent une collection. Chaque ressource est décrite par des métadonnées codées en XML (Dublin-Core / OLAC) :

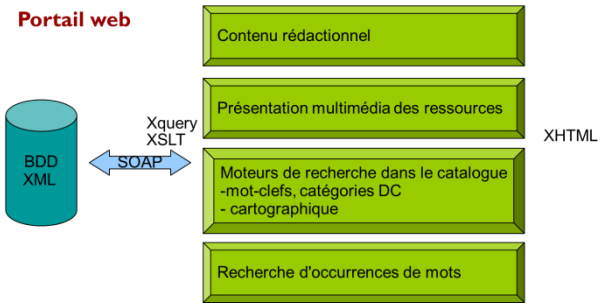


L'ensemble des métadonnées publiques de l'entrepôt est accessible via le protocole défini par l'OAI. Toutes les ressources à destination du portail forment, de manière non-exclusive, une collection « langues de France » (correspond à un *set* OAI). Enfin le moissonnage de l'archive ouverte de l'entrepôt se fait de la manière suivante :

- un set "français et langues de France".
- une metadataPrefix: olac
- tous les jours les nouveaux enregistrements ainsi que les enregistrements modifiés mettent à jour la base de données XML.



Ensuite, un portail donne accès aux corpus de cet entrepôt à partir de diverses fonctionnalités décidées dans le cadre du programme : l'ajout de contenu rédactionnel, une présentation multimédia des ressources, des moteurs de recherche sur les documents et un moteur de recherche dans les transcriptions et ou traductions.



4.4.4 Le portail Corpus de la parole [retour]

Le portail "corpus de la parole" donne accès à la collection du CRDO du même nom ainsi qu'à des informations supplémentaires sur les langues de France. On peut le décomposer en ses 3 principales composantes :

1. Le gestionnaire de contenu SPIP. Ce gestionnaire permet d'ajouter, modifier ou supprimer des contenus par le biais de formulaires web accessibles aux seuls rédacteurs autorisés. Ce gestionnaire permet de publier ces contenus sur la Toile dans une forme paramétrable.

2. La base de données XML native (eXist). Elle contient une copie de toutes les métadonnées et transcriptions.

3. Un programme Java de moissonnage (au sens de l'OAI). Ce programme récupère toutes les métadonnées de la collection « LanguesDeFrance » du CRDO et les stocke dans la base de données (eXist). Au cours du traitement, les transcriptions sont récupérées, éventuellement reformatées dans le format cible puis stockées dans la même base de données (eXist).

Dans la situation actuelle le portail SPIP est hébergé sur la grille Adonis au CC-IN2P3, mais la base de données eXist, tout comme le programme de moissonnage, reste hébergée sur COCOON en attendant que le CC-IN2P3 puisse prendre le relais. La base de données du portail contenant les métadonnées et les transcriptions sert à rassembler l'ensemble des informations en un même lieu afin d'offrir plus facilement des services (moteur de recherche, valorisations particulières). En aucun cas il ne s'agit d'une base de gestion. En particulier, le programme de moissonnage remplace les informations enregistrées par d'autres plus récentes à chaque mise à jour exécutée sur COCOON. En cas de perte ou de problème d'intégrité, il est possible de vider la base et de la reconstruire en déclenchant une opération de moissonnage complet.

Le plugin qui a été développé pour "corpus de la parole" est situé dans le répertoire ldf, lui-même compris dans le répertoire plugin du site. Il contient un script écrit en php (fichier ldf.php) qui script permet de transformer les balises particulières qui se trouveraient dans le contenu des pages du site en un code html prêt pour l'affichage. Les balises que le plugin transforme peuvent revêtir l'une des formes suivantes :

<occ_ldf/>
: Permet d'afficher un moteur de recherche sur les mots des transcriptions ou des traductions

<showtext_ldf id=oai_ID</showtext_ldf>
: Permet d'afficher la transcription correspondant à l'identifiant oai_ID

<showmeta_ldf id=oai_ID</showmeta_ldf>
: Permet d'afficher les métadonnées correspondant à l'identifiant oai_ID

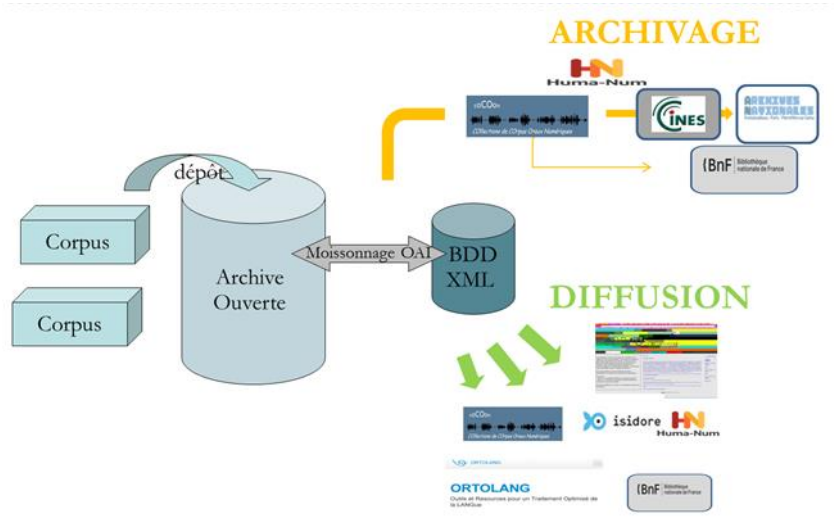
<search_ldf/>
: Permet d'afficher un moteur de recherche pour lister les ressources en fonction de critères sur les métadonnées dublin-core

<search_ldf_geo/>
: Permet d'afficher un moteur de recherche pour afficher les ressources sous forme de carte géographique

Quelle que soit la forme de la balise utilisée, une requête est exprimée (en langage XQuery) puis envoyée à la base de données (avec le protocole SOAP). Les résultats sont ensuite récupérés et transformés en HTML (par une des feuilles de style XSLT). La balise est alors remplacée par la séquence en HTML pour un affichage final de la page.
Pour des informations plus détaillées sur ce plugin, les commentaires insérés dans le code sont suffisants.
Le répertoire ldf contient également:

- un répertoire query-eXist-0.5 qui est une bibliothèque php pour l'interface SOAP de la base de données eXist,
- les feuilles des styles XSLT,
- un fichier de paramétrage "plugin.xml" qui conditionne SPIP pour l'exécution du plugin,
- quelques images utilisées pour l'affichage du code HTML,
- quelques scripts (en javaScript) utilisés pour l'affichage du code HTML.

L'architecture générale est donc la suivante :



4.4.5. Etat du *Corpus de la parole* et bilan [retour]

Au 31 juillet 2015, l'entrepôt contient 4195 documents qui se répartissent de la façon suivante :

- Espace public (données entièrement accessibles par le site Corpus de la parole) :
  - 3054 enregistrements sonores
  - 104 enregistrements vidéo
  - 440 transcriptions
  - Durée totale : 1284 heures 30 minutes

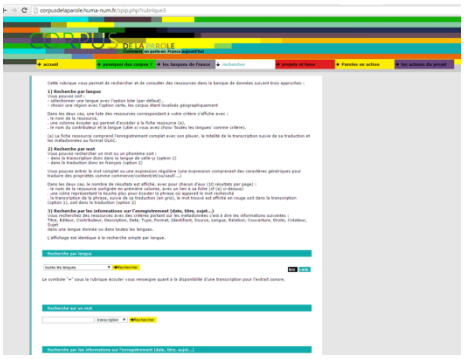
  - 49 libellés distincts de langues (correspondant à 41 codes ISO de langues distincts) : français, occitan, palikur, breton, basque, mahorais, berbère, langues kanak, créoles, judéo-espagnol, LSF, alsacien…
  - 54 déposants différents

- Espace privé (documents non accessibles en attente de correction avant validation)
  - 474 enregistrements correspondant à 737 heures
  - Documents en cours de traitement

Le site actuel contient un contenu éditorial sous la forme principalement de fiches sur les langues (exemple pour le catalan : http://corpusdelaparole.huma-num.fr/spip.php?article33).

Il permet également un accès aux corpus à l'aide des moteurs de recherche suivants :

recherche par langue,
recherche par mots,
recherche par métadonnées.



L'entrepôt abonde en corpus et en projets de toutes sortes. C'est incontestablement la plus grande réussite du programme. Les efforts et contraintes de mutualisation paraissent largement acceptables à ce niveau. Tous les corpus sont archivés au plus haut niveau de sécurité de conservation et tous disposent d'un signalement accessible aux infrastructures internationales. Le bilan est moins engageant concernant la réutilisation et la valorisation de ces corpus. C'est sur ces aspects que porte la deuxième phase du programme.

Enfin, à partir de 2010, le programme Corpus de la parole s'est heurté à de graves difficultés qui ont provoqué un fort ralentissement de son activité en 2012 et 2013.

En 2013, à la demande conjointe du conseil scientifique de l'Observatoire et des fédérations de recherche, partenaires du programme Corpus de la parole, une expertise a été réalisée par Benoit Habert et Karen Fort. Il s'agit d'un document extrêmement précieux pour tirer le bilan d'un programme expérimental qui s'est développé au fil d'un projet pilote destiné à tester l'archivage des données de la recherche. Ce document n'a pas été diffusé mais il a servi de document de travail pour élaborer la phase deux du programme qui a commencé en 2015.

Si le rapport d'expertise mérite une consultation complète, nous pouvons néanmoins en présenter les points les plus significatifs qui sont riches d'enseignements. (cf. https://www.nakala.fr/data/11280/3eddf8ef),

- **La reconnaissance de la pertinence du programme et de son rôle pionnier :**

  « Le projet Corpus de la parole a joué un rôle pionnier pour l'accord-cadre entre le CNRS et le Ministère de la Culture et pour faire avancer la diffusion, la valorisation et l'archivage pérenne des données produites par la recherche. Il est actuellement en difficulté en raison de défaillances et de changements institutionnels, dans un contexte, concernant les corpus, qui a profondément changé et qui n'offre pas aujourd'hui de garanties institutionnelles. » (Habert et Fort, Rapport 2013 :4)

- **Un projet SHS qui nécessite des compétences des sciences de l'ingénieur :**

  Le programme Corpus de la parole est un exemple de projet qui se développe avec l'apparition des « humanités numériques », bouleversant des pratiques de recherche et requérant, de la part des chercheurs travaillant sur ces programmes, des compétences spécifiques :

  « Le commanditaire (la DGLFLF) et la maîtrise d'ouvrage (les fédérations TUL et ILF) n'ont pas perçu au départ – en 2005–, et ne pouvaient d'ailleurs sans doute pas le percevoir à l'époque, le fait que Corpus de la parole constituait en définitive un projet au sens des sciences de l'ingénieur» (Habert et Fort, Rapport 2013 :4)

  « On note en outre un décalage classique entre les objectifs relativement généraux du commanditaire et de la maîtrise d'ouvrage et leur manque de connaissance/assurance quant à ce qu'il convient de réaliser. La maîtrise d'œuvre (en l'occurrence le « CRDO Paris », c'est-à-dire en définitive M. Jacobson), au lieu d'assurer la réalisation dans le cadre d'un plan fixé, se retrouve au pilotage effectif. La maîtrise d'ouvrage en est déstabilisée. Au fil du temps, les directeurs de fédération font partiellement office de

responsables de programme, ce qui n'est pas leur rôle, et sans qu'ils en aient le temps ou les compétences et le représentant de la DGLFLF prend en charge certaines des tâches d'un chef de projet. Le fait que les motivations et les intérêts des parties prenantes soient disparates n'est pas gênant en soi, mais ces divergences, si elles ne sont pas explicitées et prises en compte (comme ici), peuvent poser d'importants problèmes de compréhension et de communication. Cela a été le cas. » (Habert et Fort, Rapport 2013 :10)

- **Une défaillance de la politique en matière d'infrastructure numérique :**
  Dans le cas d'une mutation des pratiques scientifiques, l'appui institutionnel est primordial. Pour le programme Corpus de la parole, il a été salutaire au démarrage du projet puis, au gré de changements à la tête de l'infrastructure Adonis, il a été la victime de décisions préjudiciables.

  « *La deuxième phase (2010-) est marquée par les discontinuités et le multipolarisme (le recul d'un pilotage centralisé de la recherche). Les CRN ne sont plus financés par le CNRS, mais aidés plus modestement par le TGE Adonis, leur statut est incertain. En parallèle à la réorganisation du CNRS en instituts, on constate des hésitations sur la politique en matière d'infrastructure en SHS : création de l'infrastructure de recherche Corpus (et de consortiums labellisés dont l'IRCOM – Corpus oraux et multimodaux en 2011), puis fusion avec le TGE Adonis en 2013 (HUMA-NUM) ; sortie de la France du projet européen de traitement des langues (CLARIN). Parallèlement, la reconnaissance de l'Equipex ORTOLANG (2013-2019) crée un nouveau réseau d'appui concernant les ressources pour la langue française. Ce réseau rassemble des acteurs importants du domaine (ATILF, LPL, MoDyCo, LLL), mais la stratégie de cet Equipex a une autonomie relative par rapport aux efforts pour structurer des infrastructures. »* (Habert et Fort, Rapport 2013 :11)

- **Un projet qui tient grâce à l'investissement d'individus :**
  La défaillance institutionnelle provoque généralement l'arrêt du programme sauf quand des initiatives individuelles pallient ce manque. Ce ne saurait être une solution viable dans une démarche par « projets ».
  « *Dans ce contexte incertain, et de manière compensatoire, il y a eu parallèlement des fonctionnements « corsaires » : des personnes ont assuré des missions – hors institution ou dans des cadres précaires (CDD) – qui ont contribué à la réussite d'une partie des objectifs et au brouillage pour le reste. Au prix de tensions et d'incompréhensions fortes, ces personnes ont montré leur volonté tenace de faire fonctionner le projet malgré la*

disparition d'un cadre global cohérent. C'est un atout à ne pas dilapider. Il faut cependant encadrer ce travail pour éviter les « captations » mais aussi les conflits non arbitrés sur les périmètres, les responsabilités, les compétences ». (Habert et Fort, Rapport 2013 :4)

- **Une défaillance en termes de gestion de projet**
  Le rapport pointe clairement la défaillance de gestion de projet, que ce soit au niveau du pilotage ou au niveau des techniques et outils.

- **La difficulté de collecter des corpus de différents projets :**
  Le programme Corpus de la parole a mis en évidence l'extrême difficulté de la phase de collectage des corpus. C'est une phase chronophage et particulièrement délicate à mener.
  «*La recherche et l'obtention de nouveaux corpus à intégrer (investigations, contacts, négociations) ont été pesantes, à la fois parce que le repérage n'était pas évident et parce que la fourniture des corpus identifiés n'était pas toujours la priorité des chercheurs contactés.*» (Habert et Fort, Rapport 2013 :14)
  La seconde difficulté relevée est la lourdeur du travail de préparation des données :
  « *Préparation des corpus : une tâche lourde, à outiller La préparation de corpus aux fédérations et leur transmission pour versement ont été difficiles du fait de la lourdeur et de la complexité des opérations pour les fédérations par rapport aux ressources et compétences disponibles* » (Habert et Fort, Rapport 2013 :14)

- **Une réussite majeure : la procédure d'archivage :**
  « *la pérennité des données est acquise Le processus de versement et d'archivage est fonctionnel. C'est une réussite discrète mais significative, un acquis double : les données sont effectivement pérennisées (au-delà de simples sauvegardes) et les techniques et les procédures sont rodées. Le projet Corpus de la parole garde en cela sa dimension pionnière, puisque, sauf erreur, les autres archivages numériques pérennes en SHS ne portent pas sur des données de la recherche, mais sur des publications (HAL, Persée). Le processus qui s'est mis en place, sans pour autant avoir été négocié et acté entre les parties prenantes, est le suivant : COCOON (MJ) met en forme les ressources et les stocke sur la grille Adonis, puis envoie ces ressources dans l'archive du CINES. Les ressources, une fois validées au CINES comme « paquets d'archivage », sont répliquées sur le CC-IN2P3 pour diffusion. Le*

dépôt sur le TGE est manuel et s'effectue via un compte COCOON, auquel les fédérations n'ont pas accès. » (Habert et Fort, Rapport 2013 :14)

- **Un apprentissage sur le tas : la naissance d'un « projet » en humanités numériques**

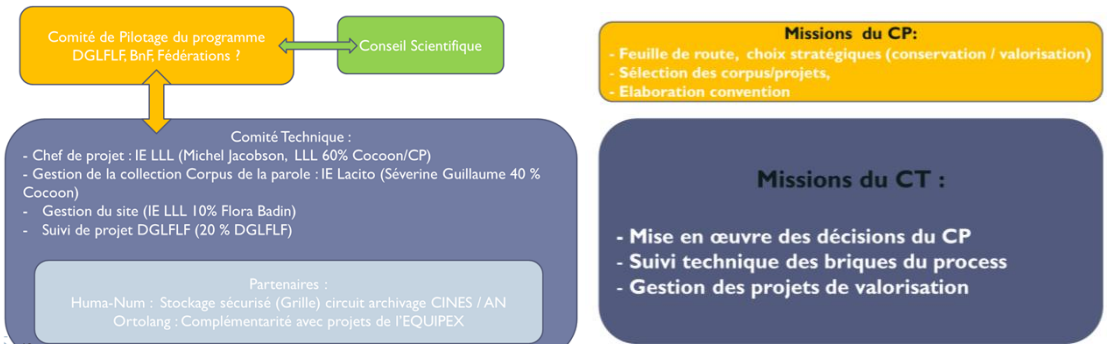
  « Le commanditaire et la maîtrise d'ouvrage n'ont pas perçu au départ le fait que Corpus de la parole constituait en définitive un projet au sens des sciences de l'ingénieur, avec une répartition maîtrise d'ouvrage/d'œuvre, avec une organisation spécifique… Ils n'ont donc pas mis en place un processus de gestion de projet : définition des objectifs du projet, des étapes de réalisation, des livrables, budget de réalisation… Rappelons que la gestion de projets (y compris dans ses déclinaisons récentes, comme la version « agile ») est un des volets fondamentaux de la formation des ingénieurs (avec de très nombreux ouvrages). Ce n'est pas un hasard si le 2e MOOC (Massive Open On-line Course) significatif lancé en France en 2013 l'a été par l'Ecole Centrale de Lille et sur ce thème. Le « mode projet » est devenu « commun » en SHS également, mais tardivement, à partir de 2007, en raison de l'évolution des modes de financement de la recherche (appels ANR), dans le vocabulaire du moins, sans que les méthodologies et les outils associés aient forcément donné lieu à appropriation réelle (on pourrait parler d'« apprentissage sur le tas »). (Habert et Fort, Rapport 2013 :16)

La lecture du rapport permet de tirer un bilan synthétique du programme. Si la quantité de données archivées prouve sa pertinence et sa faisabilité, les difficultés rencontrées sont également instructives. D'une part, il ne se présente pas de difficultés techniques majeures pour la mutualisation de corpus ; d'autre part, nous constatons la nécessité de disposer de compétences en termes de gestion de projets au sens des sciences de l'ingénieur. Toutefois la nature de l'objet scientifique et l'importance de la démarche méthodologique du chercheur en matière de corpus numérique montrent qu'il ne s'agit pas de transférer des projets des sciences humaines et sociales vers les sciences de l'ingénieur mais bien de trouver un point de rencontre dans un domaine que nous pouvons définir comme étant celui des « humanités numériques ».

Perspective : Phase 2

A partir de cette première phase du programme et du rapport subséquent, une phase 2 est en cours d'élaboration.

Elle repose sur la mise en place d'un comité de pilotage, d'un comité scientifique, d'un comité technique avec un chef de projet et d'une explicitation des tâches des parties prenantes.



Enfin un effort particulier porte sur la diffusion et la valorisation des corpus. Un nouveau site sera lancé début 2016.

Il s'appuie sur un projet innovant : « sémantisation du Corpus de la parole » qui vise à reconfigurer le portail de présentation pour tirer parti des nouvelles orientations de la Toile. En particulier, il est prévu de recourir à de l'annotation collaborative, de lier les informations de la base de données à d'autres gisements de données tels que DBpedia et à d'autres référentiels linguistiques, géographiques et d'autorité, enfin de ré-exposer ce travail de lien et d'enrichissement en suivant les principes du « Web de données ». Le portail « Corpus de la parole » qui constitue le produit le plus visible du projet du même nom a été mis au point en 2006. Initialement conçu par l'Institut de l'Information Scientifique et Technique, ce portail a peu évolué en sorte qu'il est devenu peu satisfaisant en termes d'interface et de fonctionnalités.

En particulier, il se présente comme un site web classique. L'aspect dynamique du site est principalement dû au dynamisme de la base de données qui est alimentée par le CNRS ou par les contenus éditoriaux qui peuvent être saisis par des rédacteurs de la DGLFLF et /ou du CNRS. En revanche, l'internaute (les locuteurs d'une langue ou des associations culturelles par exemples) ne peut pas ajouter directement des contenus. Il doit prendre contact avec les tutelles du projet pour faire remonter des informations qui peuvent par ailleurs être utiles à la documentation et à la diffusion des ressources présentes sur le portail. Une autre limitation est due aux technologies utilisées pour structurer et exposer les ressources. Elles ne facilitent ni le référencement ni la réutilisation des ressources et ont tendance à enfermer les données dans le seul usage prévu par l'interface du portail. L'objectif est de mieux valoriser les données du projet qui pourront être réutilisées dans d'autres contextes, de les augmenter en utilisant celles provenant d'autres sources d'information, par exemple appliquer le multilinguisme pour les concepts, et donner la possibilité aux utilisateurs de

contribuer à l'enrichissement des données. Cet objectif se décompose de manière opérationnelle en quatre sous-objectifs :

1. Décloisonner les données en les liant à d'autres gisements d'information. Un modèle de données (ontologie) devra être défini pour coder les concepts présents dans la base de données puis un alignement de ces concepts sur ceux d'autres sources (en particulier ceux de Dbpedia, de la BnF, de référentiels géographiques et linguistiques) sera effectué;

2. Ouvrir le site à la participation de communautés afin d'enrichir, corriger les métadonnées qui peuvent être lacunaires. Permettre aussi à l'utilisateur qui le souhaite d'apporter sa propre transcription d'un enregistrement ;

3. Refondre le site et son interface pour intégrer ces fonctionnalités collaboratives et mettre en œuvre la nouvelle charte graphique (définie en dehors de ce projet) ;

4. Ré-exposer l'ensemble des métadonnées sur le modèle du Web de données afin d'en faciliter la réutilisation par d'autres : en particulier dans le cadre d'un meilleur signalement dans la plateforme Isidore, proposée par Huma-Num.

4.4.6. Le linked open data appliqué à des ressources orales [BnF-EAD] [retour]

Dans le cadre de la plateforme Huma-Num, nous testons notamment à partir du corpus ESLO et du programme Corpus de la parole, les perspectives offertes par le linked data. Ce travail est piloté par Michel Jacobson, ingénieur au LLL et responsable de la plateforme COCOON.

Ce travail donne lieu à un article en cours de co-rédaction pour une soumission au CMLF 2016. Il m'a semblé opportun d'en fournir ici les éléments principaux qui reflètent pour ne pas dire résument, l'intérêt d'un travail conjoint entre linguistes et spécialistes de l'archivage de l'oral :

Le linked open data appliqué à des ressources orales
L'interopérabilité sur le plan du droit a notamment été abordée à travers le problème des licences utilisateurs. Plus récemment encore, le mouvement d'ouverture des données publiques (open-data) engagé par de nombreux gouvernements a donné lieu également à la définition d'autres licences. L'interopérabilité, a également été abordée au départ sur des aspects liés au caractère technique de l'outil informatique. C'est ainsi que les supports numériques ont connu une vague de normalisation leur permettant d'être lus plus facilement sur diverses plateformes. Mais les échanges de supports entre utilisateurs ont

massivement diminué pour laisser place à des échanges « dématérialisés » sur les réseaux. L'interopérabilité est alors essentiellement pensée au niveau du codage et du formatage des données. Sur ce domaine, on a pu observer dans les dernières années la naissance d'un certain nombre de standards et de normes très importants comme le standard Unicode pour le codage des caractères (normalisé au sein de l'ISO-10646) ou le standard eXtensible Markup Language (XML) du W3C. Sur ces briques de base, de nombreux autres formats (par exemple « Office Open XML » et « Open Document Format » pour la bureautique ; SVG pour l'image vectorielle), protocoles (par exemple SOAP, OAI-PMH) et langages (par exemple « Mathematical Markup Language » pour les expressions mathématiques, « Music Markup Language » pour la notation musicale) ont pu se construire.

Aujourd'hui, les technologies du web sémantique (le langage RDF, les ontologies RDFS, OWL, le langage de requête SPARQL…) apportent de nouveaux outils pour le codage des modèles élaborés au sein des différentes communautés. Ces technologies permettent d'expliciter formellement la sémantique contenue dans ces modèles. L'aspect formel et standardisé permettent aux machines, non pas de comprendre l'information mais de pouvoir la traiter de manière adaptée (inférences) et automatique.

Avec ces nouvelles technologies, le mouvement Open Data a pu trouver un cadre technique pour la mise en œuvre de ses principes d'ouverture et d'interopérabilité. Plus précisément le Linked Open Data pose les bases d'un modèle d'organisation intimement intégré dans l'écosystème du Web (identification des choses avec des URIs, disponibilité des données en format RDF, négociation de contenu, liage entre les entrepôts) définissant une nouvelle strate du Web appelée également Web de données ou Web 3.0.

La plateforme COCOON dédiée à l'archivage des corpus oraux, a suivi à sa création en 2006 les recommandations de la communauté OLAC (Open Language Archive Community). Ces recommandations tiennent essentiellement en deux points : l'utilisation du format OLAC pour le codage des métadonnées et l'utilisation du protocole OAI-PMH pour la diffusion de ces métadonnées. Le format de description d'OLAC est défini comme une spécialisation du Dublin-Core. La forme que prend cette spécialisation est un schéma XML classique facilitant ainsi la validation et la création d'outils d'édition.

Quelles sont les limites à ce modèle ?

Une des premières difficultés dans l'utilisation de ce modèle, mais qui représente également une de ses plus grandes forces, est sa simplicité. Cette perception de simplicité est généralement héritée de la représentation que l'on se fait en première approche du DC : « indispensable, mais jamais suffisant ». A l'analyse, le schéma n'est pourtant pas si pauvre lorsqu'on fait usage de toute sa richesse en termes de rubriques d'informations et de précision d'encodage. Ainsi, s'il n'existe effectivement qu'une seule étiquette pour les informations de localisation géographique (spatial), il est possible d'en utiliser plusieurs pour

donner plusieurs localisations, plusieurs types de localisation, ou pour exprimer une localisation en plusieurs langues. Enfin il est aussi possible d'utiliser les mécanismes d'extension du DC pour préciser les acceptions plus étroites d'une rubrique ou pour réutiliser ses propres syntaxes et vocabulaires comme le fait par exemple OLAC pour le domaine linguistique. Pour autant, si la multiplication des étiquettes de même type permet d'être plus précis, le modèle ne permet pas d'indiquer les éventuelles relations qui peuvent exister entre elles.

Une autre difficulté d'utilisation résulte du choix fait sur la plateforme COCOON de faire des descriptions au niveau des documents. Ce choix a été guidé par la volonté de décrire le plus finement possible les ressources. Il est effectivement plus facile de décrire précisément le type ou le format d'un enregistrement ou d'une transcription qu'un regroupement des deux. Ce choix est également dû au fait que le mode de production et le cycle de vie des documents n'est pas obligatoirement le même d'un type de document à un autre.

Enfin une dernière difficulté est que, pour donner aux utilisateurs une représentation intelligible des objets décrit dans la plateforme (collections, enregistrements, transcriptions, etc.), il faut composer des interfaces mélangeant des informations tirées de plusieurs ressources et qu'il est difficile pour une machine de faire la part entre les ressources, leurs relations et leurs représentations.

Afin de dépasser ces limites, la plateforme COCOON a décidé de recourir aux technologies du web sémantiques en mettant en œuvre le modèle d'exposition des données du Linked Open Data. En effet, la principale piste de progression dans la gestion des données de la plateforme COCOON est envisagée à ce jour par un changement du modèle de description. Ce changement tente de conjuguer harmonieusement la finesse du grain (au document) et la clarté du codage. Le modèle de donnée de départ est celui d'OLAC expliqué plus haut, alors que le modèle d'arrivée est fondé sur le langage RDF. Nous discuterons dans ce qui suit les choix effectués dans la mise en œuvre de ce changement.

Ce changement de modèle a suivi 3 étapes : 1) une étape d'identification des « choses » ; une étape de définition formelle du modèle cible et 3) la migration des informations de l'ancien vers le nouveau modèle.

En première étape, et comme un préalable, nous avons souhaité aligner un certain nombre de rubriques d'information sur des référentiels externes. Les objectifs initiaux étaient :

- de pouvoir identifier les « choses » plutôt que de simplement les nommer avec des littéraux. Identifier permet d'éviter l'homonymie en distinguant des individus porteurs du même nom. Le nombre de chercheurs et de locuteurs référencés dans la plateforme croissant, la présence d'homonymes augmente au risque de perturber la recherche ou du moins l'interprétation des résultats ;

- de pouvoir enrichir les informations affichées avec les informations tirées des référentiels utilisés. En effet, si la description de ces « choses » n'est pas au cœur de notre métier, nous préférons utiliser les identifiants d'autres organismes à la condition que ces identifiants soient pérennes. C'est ainsi que nous importons la description des lieux, des auteurs, des langues et des mots-clés. Cela permet par exemple de donner pour une langue ses différentes appellations, son système d'écriture, son rattachement à un arbre phylogénétique, pour un auteur sa bibliographie ou sa biographie, pour un lieu ses niveaux englobant (région, pays…), sa monnaie, son histoire....

- de limiter la redondance en évitant la duplication dans les métadonnées des ressources d'une même information en différentes langues ou à différents niveaux de précision. La récupération de ces informations pouvant être faite en interrogeant les référentiels.

Une conséquence secondaire est que ce premier travail facilitera la transition vers notre futur modèle de données cible en RDF en définissant ou en réutilisant un certain nombre d'URI. Une autre conséquence sera de pouvoir améliorer les fonctions de recherche par l'utilisation des concepts (mots-clés) dans toute leur richesse avec leurs formes génériques et spécifiques, leurs formes non préférentielles ou encore les formes équivalentes dans d'autres langues.

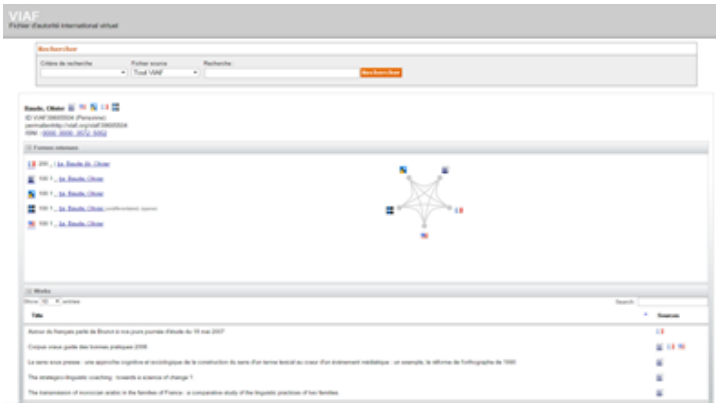
Premier exemple d'alignement : les fichiers d'autorité

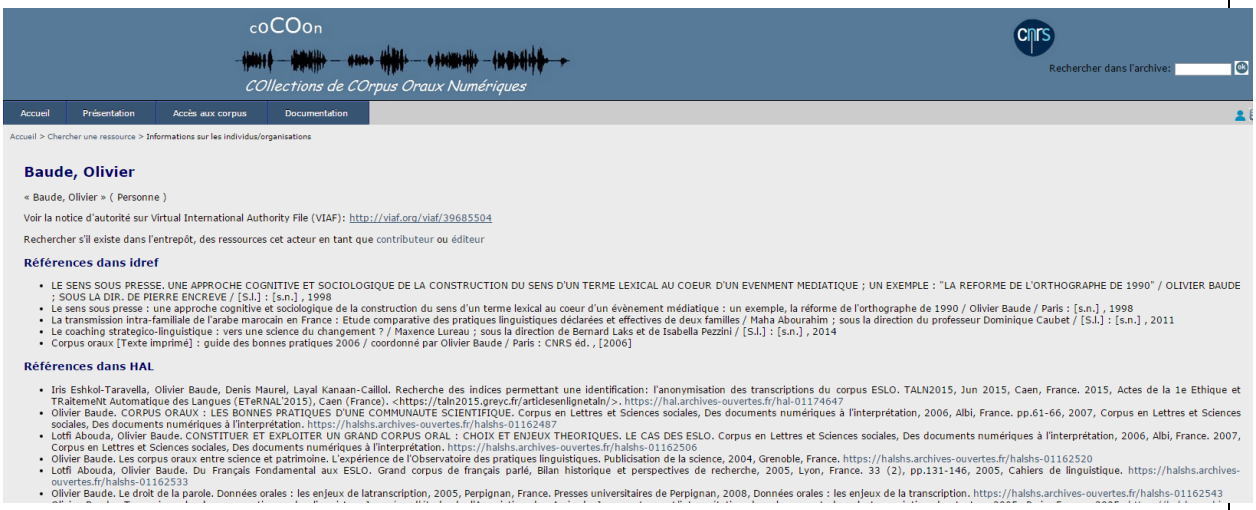
Dans un premier temps nous avons aligné une partie des contributeurs (ceux dont le rôle était « déposant ») avec le référentiel VIAF (Virtual International Authority File). Ce référentiel est né de la volonté de lier entre eux les référentiels d'autorité des bibliothèques nationales et autres grands catalogues. Si un déposant a déjà publié, il y a de fortes chances que l'on puisse trouver une ou plusieurs notices le concernant dans des bibliothèques et donc qu'un identifiant VIAF lui ait été attribué. Cet identifiant peut alors être utilisé pour identifier de manière unique l'auteur, ainsi que pour accéder à des informations le concernant. Dans la plateforme COCOON cet alignement est utilisé pour pouvoir récupérer et afficher dynamiquement des informations complémentaires. Pour chaque déposant, une page est donc construite affichant le nom normalisé ou préférentiel de l'auteur ainsi que la liste de ses publications dans les principales sources françaises (Abes, BnF). L'entrepôt HAL (Hyper Article en Ligne) a aussi été ajouté car il permet de récupérer également de la littérature grise (préprints, articles non édités, etc.) et qu'il permet éventuellement un accès direct aux documents.

La récupération des informations de la Bnf se fait par l'interrogation en SPARQL de data.bnf.fr car l'alignement avec VIAF y est explicite. La récupération des informations de l'ABES se fait en utilisant dans VIAF l'identifiant Idref de l'auteur, s'il existe, puis en retrouvant la notice correspondant à cet identifiant dans un format XML/RDF. Enfin la

récupération des informations de HAL se fait en utilisant un webService avec en paramètre l'identifiant VIAF de l'auteur. Toutefois l'introduction de cet identifiant dans HAL est très récent et demande une intervention des auteurs eux-mêmes pour le renseigner, de sorte que très peu d'auteurs l'ont pour l'instant fait.

Cet alignement va être poursuivi pour d'autres contributeurs, notamment ceux dont le rôle est « chercheur » car ils ont également de fortes chances d'avoir déjà publié en sorte qu'on puisse récupérer leurs identifiants VIAF. Pour les autres rôles, notamment les « locuteurs », nous envisageons de construire notre propre référentiel qui pourra ainsi porter en toute autonomie les propriétés de descriptions que les projets de recherche utilisent (par exemple des informations sociolinguistiques, des codes d'anonymisation, etc.).

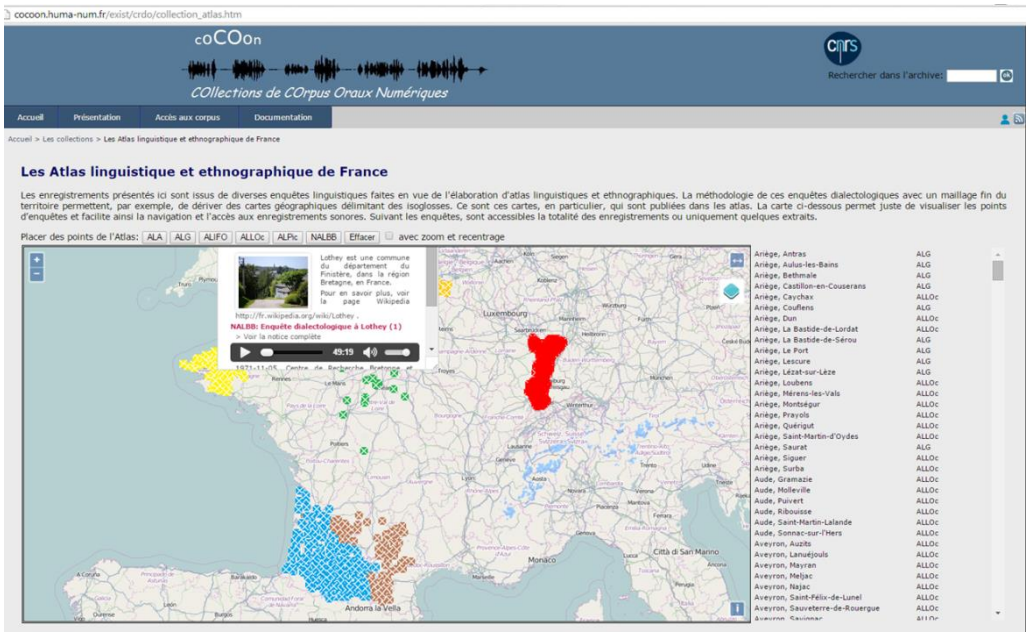




Deuxième exemple d'alignement : les lieux d'enregistrement

Pour les données audio ou vidéo de la plateforme COCOON, nous avons tenté de systématiser le renseignement du lieu géographique des enregistrements. Cette information peut représenter une donnée assez importante comme c'est par exemple le cas dans les enquêtes dialectologiques. Nous avions dans un premier temps aligné une partie de ces informations sur le TGN, puis plus récemment sur Geonames et Dbpedia. Les alignements sur Geonames ont été faits systématiquement pour les enregistrements des *Atlas*

linguistiques et ethnographiques de la France (Picardie, Bretagne, Gascogne, Alsace etc.) au niveau de la commune. Puis, à partir de l'information d'identification INSEE de la commune récupérée dans Geonames, un alignement a été fait en rebond vers le Dbpedia français. Une fois ces alignements effectués, nous avons prototypé une interface de navigation dans les enregistrements de ces *Atlas* en projetant toutes les coordonnées géographiques sur une carte de France avec une possibilité d'affichage sur chaque point des ressources disponibles dans COCOON ainsi que l'imagette de la commune, le résumé de sa présentation et le lien vers Wikipedia. Compte tenu de l'utilisation assez fréquente de ces deux référentiels ou des informations qu'ils contiennent, le rapprochement d'informations réparties dans différents gisements d'information devient une tâche beaucoup plus facile à réaliser que par le passé et il devient possible d'imaginer toutes sortes d'applications et de réutilisation à buts culturels, scientifiques ou commerciaux.



Au-delà des alignements

Les alignements ne sont pas un but en soi. Ils représentent néanmoins pour la plateforme COCOON une première étape de transition vers un nouveau mode de structuration et de mise à disposition de l'information. La deuxième étape, complémentaire, est d'identifier les « choses » qui relèvent de nos compétences ou pour lesquelles nous n'avons pas trouvé de référentiels adaptés.

Pour les ressources primaires de la plateforme COCOON (enregistrements, transcriptions, collections), nous avons déjà mis en place un système d'identification basé sur l'OAI. Ces identifiants sont repris dans des URI de type POI utilisant le système PURL . Ce sont ces POI qui seront réutilisés dans le cadre de notre futur modèle pour identifier les ressources primaires. Les autres identifiants, dont l'utilisation n'est pas systématique (les

ARK affectés par le CINES lors de leur archivage, les Handles affectés par Isidore lors de leur moissonnage) seront uniquement véhiculés dans le modèle de données comme des propriétés d'identification ou en déclarant un alignement, une équivalence entre ces identifiants et les POI.

Pour certaines « choses », nous n'avons pas encore trouvé de référentiels. Faut-il les attendre des organisations (laboratoires ou autres structures de recherche ou culturelles ayant contribué à la constitution des ressource), de l' ISNI, des locuteurs ? Il en va de même pour des objets intermédiaires comme les lieux, les évènements, la typologie.

Jacobson & Baude (soumis 2015).

4.4.7. Expériences de « valorisation » par le MCC [retour]

Les expériences les plus inattendues du programme Corpus de la parole sont celles de « valorisation » par des projets artistiques et culturels. On peut résumer les partenariats avec des artistes autour de trois lignes de force que j'ai proposé de développer au sein du programme :

- Associer la collecte à un projet artistique.

Nous l'avons souligné, une des difficultés du travail de constitution de corpus oraux est d'obtenir des données qui dépassent le paradoxe de l'observateur. Un travail en collaboration avec un projet artistique offre l'opportunité de réaliser des enregistrements dans un cadre à la fois valorisant et exempt d'enjeux sociologiques trop marqués. Participer à un projet artistique n'implique pas le même cadre primaire (au sens de Goffman) qu'une participation à un projet scientifique.

- restitution artistique et légitimité des corpus oraux

L'accessibilité à des corpus oraux structurés en base de données disponibles rend possibles des usages échappant au champ scientifique au profit d'une reconnaissance d'objets patrimoniaux au-delà de la muséographie.

- Outils numériques diversifiés

Enfin le partage d'outils de traitements des données numériques à partir de domaines distincts, permet de nouvelles approches. C'est particulièrement le cas dans le domaine de la « datavisualisation ».

Voici trois exemples :

1. Projet Trous de mémoire

(adresse du site : http://guykayser.autoportrait.com/autoportrait-collec/trous-de-memoire-le-site et journal du projet : http://gerard.paresys.free.fr/TrousDeMemoire/)

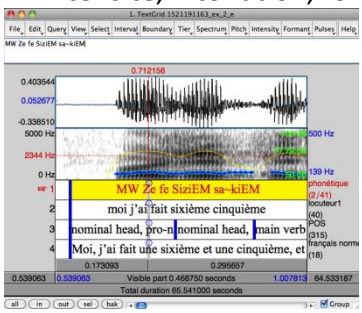
Trous de mémoire s'est construit autour de rencontres :
Rencontres entre Guykayser et les habitants du quartier Kennedy à Châlette-sur-Loing. Durant cette période vingt entretiens individuels ont été réalisés.
Rencontres orchestrées par le festival Excentrique entre Guykayser, plasticien, Olivier Baude, linguiste et Gérard Parésys, informaticien. Ce fut, tout d'abord, l'occasion de développer une thématique commune, l'«autoportrait collectif » avec des regards, des perspectives et des outils différents. Ce fut, ensuite, la réalisation d'une installation qui concrétisa l'aboutissement d'un travail fait de chemins qui se sont recoupés et entrelacés avant de se rejoindre le 22 septembre lors de la fête du quartier Kennedy.
Le travail collectif fut organisé par Guykayser entre septembre 2011 et septembre 2012. Pendant ces douze mois les rencontres s'enchaînèrent à Châlette-sur-Loing, Orléans et Paris. Il a tout d'abord fallu expliciter et confronter la démarche de chacun. Ce dialogue s'est poursuivi tout au long du projet et ce n'est pas la moindre des réussites que d'avoir créé un espace d'échanges entre artistes et universitaires.

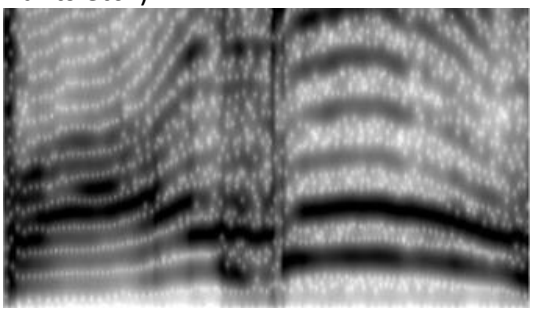
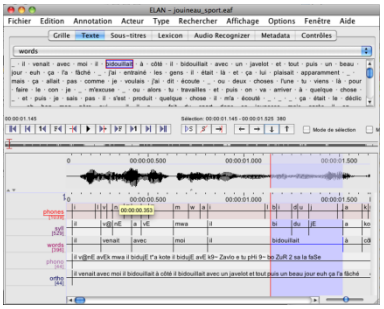
Portrait(s) et identité(s)
Dans cette perspective l'équipe du Laboratoire Ligérien de Linguistique (UMR 7270) – Olivier Baude, Emmanuelle Guerin, Céline Dugua et Caroline Cance – a proposé d'intervenir dans un premier temps à partir d'une réflexion sur les questions d'identité linguistique et de portrait sonore. C'était en effet l'occasion de confronter un projet central du laboratoire qui consiste à dresser le portrait sonore de la ville d'Orléans mais aussi de la Région Centre au travail de Guykayser. Pour cette étape, la discussion s'est appuyée sur les recherches en sociolinguistique comme le présente le texte *Portraits : regards croisés sur l'identité* rédigé par Emmanuelle Guerin et sur les réflexions méthodologiques décrites dans l'ouvrage *Corpus oraux, guide des bonnes pratiques.*
C'est parallèlement à cette réflexion que Guykayser a réalisé les entretiens avec vingt habitants du quartier Kennedy et qu'il a développé le projet concret de l'installation sonore à partir d'une image collective de vingt portraits individuels. La collaboration entre Guykayser et Gérard Parésys a permis la conception du support visuel sous la forme d'une photo de 5mX3m et du système sonore permettant l'écoute simultanée des vingt récits diffusés dans des haut-parleurs intégrés au support visuel. Celui-ci permet la présence de vingt participants dont le visage apparaît dans les trous prévus à cet effet.
Une deuxième piste de réflexion collective a été provoquée par la réalisation des montages sonores des entretiens qui, d'une heure chacun, devait donner lieu à un récit d'une à deux minutes. Sur ce point l'équipe a préféré respecter la démarche intuitive de l'artiste. Une exploitation ultérieure est prévue sous la forme d'un travail d'analyse des opérations successives du montage sonore. Quels sont les éléments conservés et quels sont les éléments systématiquement écartés du montage (on pense notamment aux disfluences, c'est-à-dire aux marques de l'oral comme les hésitations, reprises, silences) ? Quel sont les motivations explicites et implicites des choix du monteur ? Quel est le poids de la norme linguistique dans ce travail ?

Portrait(s) : des paroles en image(s)

La troisième piste de collaboration était beaucoup moins prévisible et elle s'est véritablement développée au fur et à mesure de l'avancée du projet. Le cœur de cette réflexion est la forme graphique des paroles et l'impact de la réception des variations graphiques voire orthographiques. A l'origine, il y a une étude du Laboratoire Ligérien de Linguistique sur la transcription des enregistrements sonores. Dans le cadre de ce projet, l'accent a été mis sur la diversité des formes graphiques que peut prendre une parole enregistrée. En effet, la lecture de transcriptions de paroles spontanées, transcriptions scientifiquement rigoureuses et contenant toutes les marques de l'oral, produit inexorablement un effet de stigmatisation des locuteurs. Cet effet est si puissant qu'il peut aller jusqu'à produire une autocensure des locuteurs « *Moi je ne veux pas être enregistré, je parle avec des fautes d'orthographe* ».

L'équipe a donc entrepris un travail sur la restitution graphique des enregistrements sonores. Les linguistes se sont appuyés sur les outils et les méthodes utilisés fréquemment dans le domaine des sciences du langage. Chaque enregistrement a été transcrit par un chercheur et analysé à l'aide de logiciels de traitement du signal sonore (Transcriber, Praat, Elan). La matière sonore a ainsi été découpée en « phrases », mots, syllabes, phonèmes et décrite dans ses grandes caractéristiques phono-acoustique (fréquence fondamentale, intensité, intonation, formants etc.)



Cette nouvelle matière, formée d'une description scientifique des paroles et d'une relation temporelle entre cette description et le son lui-même a ensuite été traitée par Guykayser et G. Parésys à l'aide du logiciel Processing afin de produire une image visuelle exclusivement pilotée par ces informations.

- Le spectre, vertical, en niveau de gris, défile de gauche à droite.
- La durée du défilement est égale à la durée du fichier audio.
- Les mots qui apparaissent devant/derrière, en suivant l'avancée du spectre viennent de la tire 3 « Words » de Praat.
- La position verticale de ces mots est liée à la courbe Intensity extraite de Praat.
- La couleur de ces mots est aussi liée à la courbe Intensity extraite de Praat.
- Un effet de « Blur » floute progressivement les gris du spectre et ces mots.
- A droite, de bas en haut défile point par point en blanc la courbe Pitch de Praat.
- A droite, superposée a la courbe Pitch, la tire 2 « syll » en phonétique de Praat s'écrit en bleu clair.
- La taille de ces syllabes est proportionnelle à la fréquence.
- La position de ces syllabes suit la courbe de fréquence.

En changeant très peu, on peut modifier les divers constituants de l'affichage.

Cette partie du travail tout a fait imprévue a produit des résultats prometteurs. Elle ouvre de nombreuses perspectives de collaboration ultérieures. C'est en effet une contribution singulière aux travaux sur la visualisation des données. La réflexion conjointe d'artistes et d'universitaires se concrétisera par la poursuite d'installations artistiques et de projets culturels mais aussi par la production de conférences et d'articles scientifiques.

L'installation *Trous de mémoire* le 22 septembre

[Le dispositif a été mis en place quelques jours avant le jour de la manifestation ce qui a permis un « vernissage » en présence des vingt participants.]

L'espace

L'espace de l'installation a été divisé en trois grandes parties.

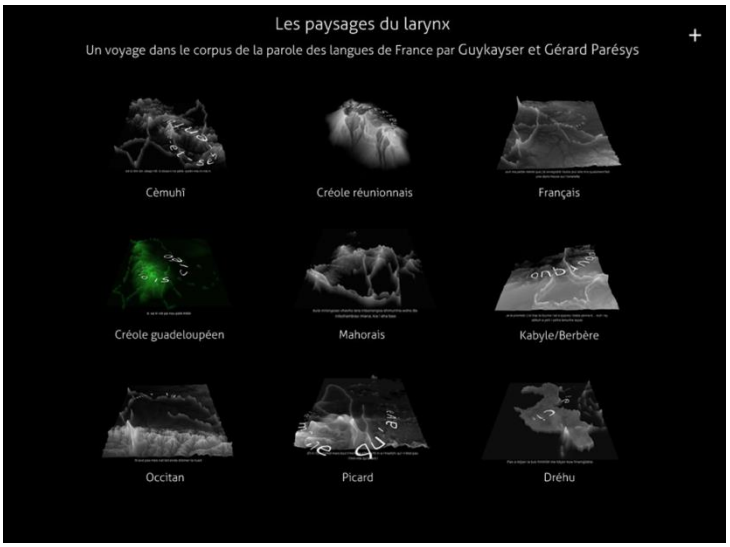
La première accueillait les participants qui prenaient place derrière la photo et écoutaient le portrait sonore du personnage choisi.

La deuxième partie était consacrée à un espace technique : prise de vue, poste informatique pour le traitement de l'image, impression des photos.

La troisième partie donnait la possibilité aux participants de s'assoir, d'écouter tous les portraits sonores en regardant un écran où défilaient les images issues du traitement graphique et plastique de l'enregistrement.

2. Les paysages du Larynx

Les paysages du Larynx est une œuvre numérique réalisée par Guykayser et Gérard Parèsys en collaboration avec Olivier Baude à partir des données du site Corpus de la parole. Reprenant le travail fait sous Praat et processing, des données de géolocalisation étaient ajoutées à la transcription et aux informations sur le signal pour générer une cartographie artistique à partir de corpus en langues de France. Cette œuvre sera intégrée au nouveau site en 2016.



3. Le cabinet de curiosités des langues de France

Le cabinet de curiosité est un projet réalisé par l'association Labomédia et le LLL. Il a été sélectionné dans le cadre de l'appel à projet « services culturels innovants » du Ministère de la Culture et Communication en 2014. Il donnera lieu à une première exposition fin 2015.

Présentation du projet
Le projet de « cabinet de curiosités des langues de France » est constitué d'un ensemble d'objets et de dispositifs interactifs qui donne à voir et à entendre à un large public un travail de recherche sur la langue française, de façon artistique, anachronique et participative.
Ce projet, porté par l'association Labomedia, s'appuie sur le programme Corpus de la parole (DGLFLF/MCC-CNRS) pour les langues de France et sur les enquêtes sociolinguistiques menées à Orléans (« ELSO 1, 2 et 3.0 ») par le Laboratoire Ligérien de Linguistique UMR 7270 BnF-CNRS-Universités Orléans & Tours. Il fédère ainsi des acteurs tant dans le domaine de la recherche que dans le champ de la création numérique de dispositifs culturels de médiation en impliquant dans sa réalisation un certain nombre de structures à l'échelle du territoire français, localement des groupes de personnes.
Tel qu'imaginé aujourd'hui, il se matérialise par des dispositifs interactifs aux formes multiples, voulus didactiques et ludiques, et une partie web pour parcourir de façon visuelle et sonore les corpus oraux. Il invite au croisement des regards entre artistes, scientifiques et grand public en permettant à chacun d'être acteur du projet afin de mieux appréhender les enjeux qui résident dans le langage, de sa nature symbolique à son traitement automatique comme système de représentation du monde.

Un processus de recherche-création-fabrication en réseau
Pour refléter la diversité de cet objet multifacette, ce cabinet de curiosité des langues de France s'articule autour de modules interactifs et d'objets, d'éléments plus "low-tech", avec l'envie d'impliquer des jeunes et moins jeunes dans un processus de recherche-création fabrication autour des modules envisagés et à partir d'idées originales, comme une autre façon se confronter aux corpus, de les explorer.
Ce processus de recherche-création-fabrication est aussi mis en œuvre au sein du réseau de partenaires de la recherche et du développement, dans une approche transdisciplinaire ouverte afin de favoriser les croisements, les collaborations, les mutualisations de moyens et de compétences, la sérendipité. Souhaitant instaurer au sein de ce réseau une dynamique d'échange, de partage, nous fondons la coopération sur les mêmes principes que les communautés autour des matériels et logiciels libres, de l'open source, des expériences passées laissant à penser que l'écosystème qui se développe ainsi est susceptible de produire des résultats plus sapides et variés. C'est donc un réseau de personnes et de structures que nous souhaitons densifier, ramifier à travers ce projet, réseau préexistant parfois, en perpétuelle mutation au fil des connexions et interconnexions que chacun tisse.

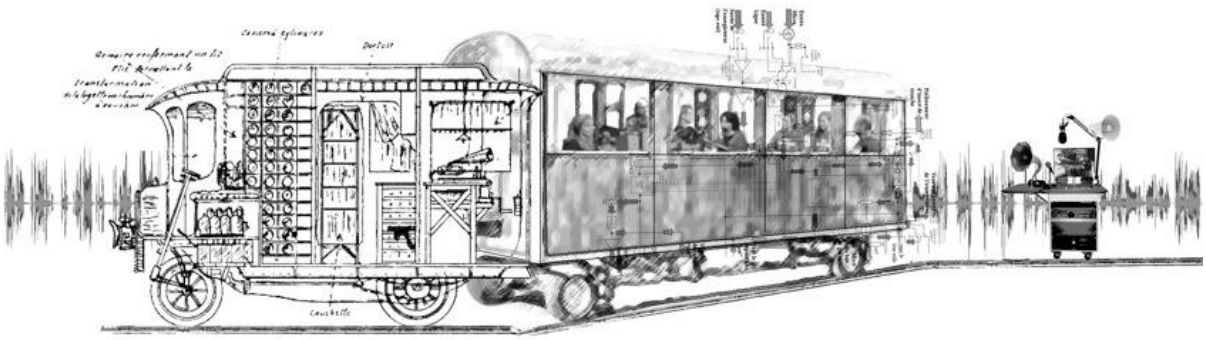
Ces différents projets ne sont que des premières expériences balbutiantes. Elles assurent toutefois un ancrage important dans le tissu social et transforment l'ensemble des opérations de la collecte à la restitution. Elles interrogent aussi d'une manière singulière la démarche du chercheur à toutes les étapes de la constitution et de l'exploitation du corpus. Cette phase transforme elle aussi l'objet scientifique lui-même.

5. Conclusion [retour]

Cette synthèse s'est proposé de donner les éléments de présentation mais aussi d'analyse réflexive d'un parcours de recherche « atypique ». Ce parcours n'est pas arrivé à son terme. Il était une étape nécessaire à la poursuite d'un programme de travail dont, au moment de mettre un point provisoirement final, force est de constater qu'il n'aurait pas été possible d'indiquer au commencement l'orientation qu'il a prise.

A titre d'ouverture, on exposera quelques perspectives de recherches dont la dernière, qui requerra un peu d'indulgence du lecteur, résume, à différents titres, certains des éléments explicités dans les pages qui précèdent. On pourra la lire comme une métaphore d'un programme scientifique.

La roulote des Corpus de la parole



COUPE SCHÉMATIQUE

Le projet de *Roulotte des corpus de la parole* est un clin d'œil, révérence gardée, à la roulotte de Ferdinand Brunot reproduite dans la partie gauche du croquis. F. Brunot avait dessiné cet équipement qui devait permettre de réaliser sur le terrain, en France, des enregistrements des parlers. La partie gauche est aménagée avec un casier où sont stockés les rouleaux enregistrés, la partie centrale est réservée au chercheur et la partie droite est une chambre d'enregistrement. Ainsi équipé, F. Brunot envisageait d'aller à la rencontre des variétés linguistiques et de leurs locuteurs afin d'en conserver le témoignage et de produire les données dont la linguistique manquait tout en élevant l'oral au rang des savoirs dignes d'un traitement documentaire et archivistique. Vaste ambition où se rejoignaient soutien académique, développement technologique et initiative personnelle.

Un siècle plus tard, l'esprit dans lequel ce projet avait été conçu correspond toujours à une motivation scientifique et sociale. Les conditions de réalisation ont changé. Si le volontarisme individuel d'un homme providentiel ne peut apporter la solution, et que le soutien universitaire reste marqué par l'insuffisance de légitimité de l'objet, l'opportunité qu'offrent des avancées technologiques maitrisées par le scientifique est bien réelle. Aussi la « nouvelle » roulote répond-elle à un double objectif : d'une part La *Roulote*, parce qu'elle crée un évènement sur le terrain, amène à concevoir la démarche d'enquête dans toutes ses dimensions et son déroulement, d'autre part, elle représente une possibilité de conjoindre collecte et restitution dans la même approche. Une captation de la variation linguistique est assurée dans la même démarche qui en assure la reconnaissance dans l'espace social et culturel.

Cet excursus en forme d'anecdote ne doit pas distraire l'attention d'un travail plus académique mais tout aussi prototypique et qui concerne l'évolution du corpus ESLO.

Après ESLO1 qui a bénéficié du développement de la sociolinguistique et de technologies de captation sonore mobiles et après ESLO2 qui s'est inspirée des avancées de la sociopragmatique et de la linguistique de corpus, il parait possible de contribuer à un nouveau tournant théorique et épistémologique à partir d'une nouvelle enquête sociolinguistique. Il s'agit moins de produire une méthodologie ou un paradigme scientifique révolutionnaires que de marquer une scansion dans un rythme afin de déployer des analyses linguistiques à partir de corpus « maitrisés ». Les éléments pour une science du corpus sont seulement esquissés dans cette synthèse mais on peut d'ores et déjà entrevoir leurs effets et la perspective qu'ils dessinent.

En partant de la démarche d'une sociolinguistique qui s'ancre dans les pratiques des humanités numériques et dans le cadre plus large du Web sémantique, plusieurs pistes de réflexion s'offrent à nous.

La première concerne la nécessité de penser les données sur l'ensemble de la chaine de « traitements » en intégrant leurs conditions de production, c'est-à-dire :

- la définition d'un objet scientifique par le chercheur. Cette « objectivation » concerne l'épistémologie du champ, l'évaluation des théories, la construction des terrains et la mesure des effets des traitements. Cela nécessite de documenter rigoureusement et en détail la démarche du chercheur et de concevoir les outils permettant de structurer et d'incorporer dans l'objet d'étude la documentation.

- La collecte, qui ne peut se soustraire à la méthodologie et aux questionnements réflexifs de l'enquête scientifique. L'enquête sociolinguistique est le premier geste de la démarche du chercheur en linguistique.

- La prise en charge des aspects juridiques et éthiques à toutes les étapes du projet de recherche et la nécessité d'expliciter la démarche dans la définition même de l'objet.

- Le « figement des données » lors des opérations de stockage et d'exposition de celles-ci. Sur ce point les linguistes ont beaucoup à apprendre d'autres sciences, en particulier les sciences documentaires. En effet, poser la question de l'interopérabilité des données oblige à prendre en compte la signification des opérations de codage et de catégorisation, de partage de vocabulaires contrôlés, de référentiels et d'ontologies et au-delà de la relation aux données. Dans le cadre de la linguistique, c'est là que la conceptualisation et la concrétisation des « données situées » se décide.

- Les traitements qui vont transformer une source, des données brutes, primaires, et les orienter en fonction de l'analyse qui leur sera appliquée.

- Le signalement des données. Devenues un objet « social », elle participent alors d'une autre dimension du savoir, celle d'objets partagés offerts à différents points de vue et à différents usages. Il est dès lors possible d'effectuer un retour sur les données liées à une analyse et leur mise à l'épreuve, d'ouvrir un dialogue à l'intérieur de la discipline à partir de la comparaison des données et une confrontation inter- et transdisciplinaire afin de reconsidérer des ressources qui se prêtent à une exploitation sociale, politique et culturelle, à des applications.

- La conservation et l'archivage qui modifient l'objet en modifiant son statut. Conserver et archiver apportent une reconnaissance des données non seulement par le chercheur mais aussi par l'institution sociale. Pour ce qui concerne les corpus oraux, cela vient renforcer le lien entre le monde de la recherche et celui de la culture.

Une deuxième piste, illustrée par cette synthèse, permet de concevoir comment ces différentes étapes sont à la fois successives et imbriquées. Si elles sont souvent réalisées suivant un axe chronologique, elles sont néanmoins solidarisées par un effet de *bootstrapping*. Par exemple, poser la question de l'archivage au terme du processus est à la fois source d'échec et cause de post catégorisation.

Aujourd'hui, l'une des visions les plus stimulantes vient de l'évolution de la connaissance à partir de son architecture sur la Toile. Avec le web sémantique, la frontière entre données et métadonnées s'estompe au profit d'une nouvelle gestion des connaissances : les métadonnées deviennent des données dans un processus fascinant de génération. Nombre de travaux abordent cette question. Je me contenterai de donner un aperçu sur ce qui pourrait orienter mon programme de recherche dans les années à venir. Dans ma thèse, je me référais à la théorie de l'intégration conceptuelle de Fauconnier & Turner (2003)[147] en appliquant leurs propositions à l'usage de métaphores dans la « langue en pratique » : le fait de relier deux espaces mentaux à l'aide de compétences cognitives et linguistiques de différents niveaux produit, comme dans la métaphore, un sens nouveau qui va au-delà de ce qui résulterait d'une simple addition du sens des deux espaces initiaux. Ce processus

[147] FAUCONNIER, G., & TURNER, M. (2003). *The way we think: conceptual blending and the mind's hidden complexities*.

comprend trois étapes : la composition, la complémentation et l'élaboration. Les deux espaces sont liés au moyen de certains de leurs éléments, puis des inférences complètent un espace mental « mélangé » (blend) et permettent de générer un nouveau sens. Or, pour Fauconnier & Turner, cette faculté, qui s'exemplifie dans le langage, est une faculté plus générale du fonctionnement cognitif humain qui produit une évolution des capacités cognitives. Sans vouloir faire une métaphore sur une théorie de la métaphore, le Web sémantique ou Web de données peut être conçu comme un vaste domaine de la connaissance humaine au cœur duquel le « liage » de données construit de nouvelles connaissances. Les travaux actuels sur l'enrichissement des données montrent le potentiel de la structuration de l'architecture d'une connaissance à la fois commune et individuelle.

La troisième piste est elle aussi congruente à la perspective cognitive. Il s'agit de l'apport de la datavisualisation pour traiter à la fois des données massives et des données structurées et liées. Si une architecture de données liées, au niveau du Web, est à même de générer une connaissance nouvelle, la question de la restitution de celle-ci se pose. Les moyens élaborés par la datavisusalisation devraient contribuer à repérer, restituer et figurer ce qui n'apparait pas analogiquement. D'où l'importance de traitements multivariés des données. Le fait que celles-ci soient appréhendées par une strate iconique est au centre d'une conception de la linguistique cognitive qui ne sépare pas le langage de la perception sensorielle variée.

Ces trois perspectives impliquent de parcourir le long et difficile chemin de constitution des données en intégrant leurs conditions de production.

Pour conclure, je préciserai que la forme choisie pour exposer mon parcours de recherche se veut également une contribution aux nouvelles pratiques liées au bouleversement général de la construction et de la circulation du savoir à l'ère du numérique.

Un effort constant a été fait pour référer la présentation à la production scientifique disponible en matière de consultation et d'archivage et pour assimiler les apports que l'interopérabilité des données offrent à des recherches disciplinaires et interdisciplinaires.

Tout ceci doit être abordé dans une approche réflexive déterminante sur les conditions de production des données comme élément premier d'une science de l'observation. *[retour]*

Bibiographie

Bibliographie :

AIJMER, K., & RÜHLEMANN, C. (Éd.). (2015). *Corpus pragmatics : a handbook*. Cambridge, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press,.

ALEXANDER, J. (2004). *Frequency, Prosody, and French Liaison : Testing Bybee's Hypothesis* (Distinction in Linguistics). Boston.

AMACKER, R. (1975). *Linguistique saussurienne*. Genève, Paris : Librairie Droz.

ASHBY, W. (1981). « French liaison as a sociolinguistic phenomenon ». In W. W. Cressey & D. J. Napoli (éd.), *Linguistics Symposium on Romance Languages (9th)* (p. 46-57). Washington, DC: Georgetown University Press.

BAUDE, O. (1998). *Le sens sous presse: une approche cognitive et sociologique de la construction du sens d'un terme lexical au coeur d'un évènement médiatique* (Thèse de doctorat). Paris, France.

BAUDE, O. (2005). « Le droit de la parole ». In M. Bilger (coord.), *Données orales : les enjeux de la transcription*. Perpignan, France: Presses Universitaires de Perpignan.

BAUDE, O. (2007a). « Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux ». *Revue française de linguistique appliquée*, *XII*(1), 85-98.

BAUDE, O., MARCHELLO-NIZIA, C., MONDADA, L., BLANCHE-BENVENISTE, C., CALAS, M.-F., CAPPEAU, P., CORDEREIX, P., LAMBERTERIE, I. D., GOURY, L., & JACOBSON, M. (2006). *Corpus oraux : guide des bonnes pratiques* (O. Baude, éd.). CNRS Éditions et Presses universitaires d'Orléans.

BERGOUNIOUX, G. (1992). *Enquêtes, corpus et témoins en France, hier et aujourd'hui*. Paris, France: Larousse.

BERTRAND, O., DEPECKER, L., & PRUVOST, J. (Éd.). (2013). *Linguistique: traductions et terminologie*. Paris, France: H. Champion.

BIBER, D. (1991). *Variation across speech and writing*. Cambridge England, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press.

BIBER, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam, Pays-Bas, Etats-Unis d'Amérique : John Benjamins.

BIBER, D., & REPPEN, R. (Éd.). (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord: Cambridge University Press.

BLANCHE-BENVENISTE, C., & JEANJEAN, C. (1987). *Le français parlé: transcription et édition*. Paris, France : Didier.

BONAMI, O., & BOYE, G. (2003). « La nature morphologique des allomorphies conditionnées : Les formes de liaison des adjectifs en français ». *Silexicales*, *3*, 39-48.

BOOIJ, G., & DE JONG, D. (1987). « The domain of liaison: theories and data ». *Linguistics*, *25*, 1005-1025.

BOURDIEU, P. (1980). *Le sens pratique*. Paris: Éditions de Minuit.

BOURDIEU, P. (2015). *Sur l'Etat : Cours au Collège de France*. Paris: Points.

BYBEE, J. (2001). « Frequency effects on French liaison ». In J. Bybee & P. Hopper (éd.), *Frequency and the emergence of linguistic structure* (p. 337-359). Amsterdam, Philadelphia: John Benjamins Publishing Company.

CAMERON, D. (1992). *Researching language: issues of power and method*. London, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Routledge

CAPPEAU, P., & GADET, F. (2007). « Où en sont les corpus de français parlé ». *Revue française de linguistique appliquée*, *Vol. XII*(1), 129-133.

CHABANAL, D., & LIEGEOIS, L. « Production de liaisons dans l'input parental ». In C. Soum-Favaro, A. Coquillon, & J.-P. Chevrot (éd.), *La liaison : approches contemporaines*. Bern : Peter Lang.

CHEVROT, J.-P., CHABANAL, D., & DUGUA, C. (2007). « Pour un modèle de l'acquisition des liaisons basé sur l'usage: trois études de cas ». *Journal of French Language Studies*, *17*, 103-128.

CHEVROT, J.-P., DUGUA, C., & FAYOL, (2005). « Liaison et formation des mots en français: un scénario développemental ». *Langages*, *158*, 38-52.

CHEVROT, J.-P., DUGUA, C., & FAYOL, M. (2009). « Liaison, word segmentation and construction in French: a usage-based account ». *Journal of Child Language*, *36*(3), 557-596.

CHEVROT, J.-P., & FAYOL, M. (2001). « Acquisition of French liaison and related child errors ». In M. Almgren, A. Barrena, M. J. Ezeizabarrena, I. Idiazabal, & B. MacWhinney (éd.), *Research on Child Language Acquisition: Proceedings of the 8th Conference of the International Association for the Study of Child Language* (Vol. 2, p. 760-774).

CHEVROT, J.-P., FAYOL, M., & LAKS, B. (2005). « La liaison: de la phonologie à la cognition ». *Langages*, *158*, 3-7.

CHOMSKY, N. (1964). *Current issues in linguistic theory*. The Hague: Mouton.

CONEIN, B. (1994, Février). *La représentation des catégories sociales : taxinomie et classification* (HDR). EHESS, Paris.

CONRAD, S., & BIBER, D. (Éd.). (2001). *Variation in English: multi-dimensional studies*. Harlow, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord Routledge.

COTE, M.-H. (2002). « Between phonology and the lexicon: French liaison revisited » Conférence à l'Université de Toronto.

COTE, M.-H. (2005). « Le statut lexical des consonnes de liaison ». *Langages*, *158*, 66-78.

CÔTÉ, M.-H. (2007). « Empty elements in schwa, liaison and h-aspiré: the French holy trinity reconsidered ». In J. M. Hartmann, V. Hegedüs, & H. van Riemsdijk (éd.), *Sounds of silence: empty elements in syntax and phonology*. Oxford : Elsevier.

COTTOUR, C., REGIMBEAU, G., & CORDEREIX, P. (2008). *Méthodologie de la prospection au dépôt légal son à la Bibliothèque nationale de France* (Projet professionnel personnel de bibliothécaire : gestion de projet : bibliothéconomie). Villeurbanne, ENSSIB. .

DAUTRICOURT, R. G. (2010). *French liaison: Linguistic and sociolinguistic influences on speech perception* (Dissertation). The Ohio State University Editor.

DAVIS, J. L. (2000). *French liaison: a case study of the syntax/phonology interface* (Ph. D. Dissertation). Indiana University Editor, Bloomington.

DE JONG, D. (1988). *Sociolinguistic aspects of French liaison*. Vrije Universiteit Amsterdam Editor, Amsterdam De Gruyter.

DE JONG, D. (1990). « The syntax-phonology interface and French liaison ». *Linguistics*, *28*, 57-88.

DE JONG, D. (1991). « La liaison à Orléans (France) et à Montréal (Quebec) ». In *Actes du XIIe Congrès International des Sciences Phonétiques* (p. 198-201). Université de Provence.

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

DELATTRE, P. (1947). « La liaison en français, tendances et classification ». *The French Review*, *21*, 148-157.

DELATTRE, P. (1955). « Les facteurs de la liaison facultative en français ». *The French Review*, *29*, 42-49.

DELATTRE, P. (1956). « La fréquence des liaisons facultatives en français ». *The French Review*, *30*, 48-54.

DEPECKER, L. (2005). *La terminologie: nature et enjeux*. Paris, France : Larousse.

DEPECKER, L. (2009). *Comprendre Saussure: d'après les manuscrits*. Paris, France : A. Colin.

DESROCHERS, R. (1994). « Les liaisons dangereuses : le statut équivoque des erreurs de liaison ». *Linguisticae Investigationes*, *XVIII:* 243-284.

DESROSIERES, A. (2008). *L'argument statistique*. Paris, France: Mines Paris-Tech-les Presses,

DETEY, S., DURAND, J., & LAKS, B. (Éd.). (2010). *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement*. Paris, France: Éd. Ophrys.

DUGUA, C. (2005). « De la liaison à la formation du lexique chez les jeunes enfants francophones ». *Le Langage et l'Homme*, *40*(2), 163-182.

DUGUA, C., & BACLESSE, M. « Incidence d'effets de fréquence sur l'usage de la liaison en lecture à haute voix et dans des jugements normatifs chez des enfants de CE2-CM1 ». In C. Soum-Favaro, A. Coquillon, & J.-P. Chevrot (éd.). *La liaison : approches contemporaines.* Bern : Peter Lang.

DUGUA, C., SPINELLI, E., CHEVROT, J.-P., & FAYOL, M. (2009). « Usage-based account of the acquisition of liaison: evidence from sensitivity to plural/singular orientation of nouns ». *Journal of Experimental Child Psychology*, *102*, 342-350.

DURAND, J., LAKS, B., CALDERONE, B., & TCHOBANOV, A. (2011). « Que savons-nous de la liaison aujourd'hui ». *Langue française, 169*, 103-135.

DURAND, J., & LYCHE, C. (2008). « French liaison in the light of corpus data ». *Journal of French Language Studies*, *18*, 33-66.

ENCREVE, P. (1983a). « La « liaison » entre la linguistique et la sociolinguistique 1/2 ». *Noroit*, *280*, 2-23.

ENCREVE, P. (1983b). « La liaison sans enchaînement ». *Actes de la recherche en sciences sociales*, *46*(1), 39-66.

ENCREVE, P. (1984). « La « liaison » entre la linguistique et la sociologie 2/2 ». *Noroit*, *281*, 19-23.

ENCREVE, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français*. Paris: Edition du Seuil.

ENCREVE, P. (2002). « La langue de la république ». *Pouvoirs*, *100*, 123-136.

ENCREVE, P., FORNEL, M. de, LAKS, B., BOURDIEU, P., & LABOV, W. (1983). *Actes de la recherche en sciences sociales. L'usage de la parole,* Paris, France: Maison des sciences de l'homme.

FAUCONNIER, G., & TURNER, M. (2003). *The way we think: conceptual blending and the mind's hidden complexities*. New York, Etats-Unis d'Amérique: Basic Books.

FETEKE, J.-D. (s. d.). « The InfoVis Toolkit ». Consulté à l'adresse https://www.lri.fr/~fekete/ps/ivtk-04.pdf

FILHON, A., DEAUVIEAU, J., DE VERDALLE, L., PELAGE, A., POULLAOUEC, T., BROUSSE, C., MESPOULET, M., & SZTANDAR-SZTANDERSKA, K. (2013). « Un projet de nomenclature socioprofessionnelle européenne ». *Sociologie*, *n°4, vol. 4*.

FOUGERON, C., & DELAIS-ROUSSARIE, E. (2004). « Liaisons et enchaînements : « Fais_en à Fez_en parlant » ». In *Actes des Journées d'Etudes sur la Parole 2004*, 221-224.

FOUGERON, C., GOLDMAN, J.-P., & FRAUENFELDER, U. (2001). « Liaison and schwa deletion in French: an effect of lexical frequency and competition ». In *Proceedings of the 7th European Conference on Speech Communication and Technology, Eurospeech 2001*, 639-642.

GADET, F. (1987). *Saussure, une science de la langue*. Paris, France : Presses Universitaires de France.

GADET, F. (1997). *Le français populaire*. Paris, France : Presses Universitaires de France.

GADET, F. (2007). *La variation sociale en français*. Paris, France : Ophrys.

GADET, F. (2012). *Cahiers de linguistique (Courtil-Wodon), Construction des connaissances sociolinguistique*. Cortil-Wodon (Belgique), Belgique: E.M.E..

GOUGENHEIM, G. (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris, France : Didier.

GRECO, L., MONDADA, L., & RENAUD, P. (Éd.). (2014). *Identités en interaction*. Limoges, France : Lambert-Lucas.

GREIDANUS, T. (1990). *Les constructions verbales en français parlé : étude quantitative et descriptive de la syntaxe des 250 verbes les plus fréquents*. Tübingen, Allemagne : M. Niemeyer.

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

HABERT, B. (2009). *Construire des bases de données pour le français* : Ophrys

HABERT, B. (2012). « L'archivage numérique entre us et abus de la mémoire numérique » (p. 23-43). Présenté aux 11èmes Journées internationales d'analyse statistique des données textuelles (JADT). Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00991517.

HABERT, B., NAZARENKO, A., & SALEM, A. (1997). *Les linguistiques de corpus*. Paris : A. Colin.

HADDINGTON, P., KEISANEN, T., & MONDADA, L. (Éd.). (2014). *Multiactivity in social interaction: beyond multitasking*. Amsterdam, Pays-Bas, Etats-Unis d'Amérique John Benjamins.

HADDINGTON, P., MONDADA, L., & NEVILE, M. (Éd.). (2013). *Interaction and mobility: language and the body in motion*. Berlin, Allemagne, Etats-Unis d'Amérique : De Gruyter.

HEIDEN, S., LAFON, P., ILLOUZ, G., HABERT, B., FLEURY, S., & FOLCH, H. (1999). « Maîtriser les déluges de données hétérogènes ». In Condamines, A., ;Fabre, C.;Péry-Woodley, M. P. (éd.). (p. 37-46). Cargèse, Italy: ELDA. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00151841

HEIDEN, S., LAFON, P., ILLOUZ, G., HABERT, B., FLEURY, S., FOLCH, H., & PRÉVOST, S. (2000). « Prendre Le Monde en main : choix d'architecture ». In *RIAO 2000*. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00151840

HEIDEN, S., PREVOST, S., HABERT, B., FOLCH, H., FLEURY, S., ILLOUZ, G., LAFON, P., & NIOCHE, J. (2000). « TyPTex : Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation ». In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (éds) *Second International Conference on Language Resources and Evaluation* (p. 141-148). Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00087993

HEIDEN, S., PREVOST, S., HABERT, B., ILLOUZ, G., LAFON, P., FLEURY, S., & FOLCH, H. (2000). « Profilage de textes : un cadre de travail et une expérience ». In *JADT'2000*. Lausanne, Switzerland: JADT. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00151839

HERAN, F. (2002). « Les langues et la statistique publique, des comptages du Second Empire au volet linguistique de l'enquête famille ». *Ville Ecole Intégration Enjeux*, 13I0, 51-74.

« International Association of Sound and Audiovisual Archives ». (s. d.). Consulté 18 juillet 2015, à l'adresse http://www.iasa-web.org/

KLAUSENBERGER, J. (1974). « Rule inversion, opacity, conspiracies: French liaison and elision ». *Lingua*, *34*, 167-179.

KNORR-CETINA, K., & CICOUREL, A. V. (Éd.). (1981). *Advances in social theory and methodology: toward an integration of micro- and macro-sociologies*. Boston, Etats-Unis d'Amérique : Routledge Library Editions.

LABOV, W. (1972). *Sociolinguistic patterns*. Philadelphia, Etats-Unis d'Amérique: University of Pennsylvania Press.

LABOV, W. (1978). *Sociolinguistic patterns*. Oxford, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : B. Blackwell.

LABOV, W. (1981). *Sociolinguistic Patterns*. Philadelphia, Etats-Unis d'Amérique: University of Pennsylvannnia Press.

LABOV, W. (1993). *Le parler ordinaire : la langue dans les ghettos noirs des États-Unis*. (A. Kihm, Trad.). Paris, France: les Éditions de Minuit.

LABOV, W. (2006). *The social stratification of English in New York city*. Cambridge, Etats-Unis d'Amérique : Cambridge University Press.

LABOV, W. (2010). *Principles of linguistic change*. Oxford, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, Etats-Unis d'Amérique : Cambridge University Press.

LABOV, W. (2013). *Linguistics, ISSN 0024-3949.*. In J.-P. Chevrot & P. Foulkes (Ed.) *Language acquisition and sociolinguistic variation*. Berlin, Allemagne, Etats-Unis d'Amérique : De Gruyter.

LAKS, B. (1996). *Langage et cognition : l'approche connexionniste*. Paris, France : Hermès.

LAKS, B. (2000). *Différenciation linguistique et différenciation sociale : quelques problèmes de sociolinguistique française* (Thèse de 3e cycle). Paris.

LAKS, B. (2005). « La liaison et l'illusion ». *Langages*, *158*, 101-125.

LAKS, B. (2006). « Phonologie et construction syntaxique : la liaison, un test de cohésion et de figement syntaxique ». *LINX*, *53*, 155-172.

LAKS, B. (2008). « Pour une phonologie de corpus ». *Journal of French Language Studies*, *18*(01), 3-32.

LAKS, B. (Éd.). (2011). *Langue française (Paris. 1969). Phonologie du français contemporain*. Paris, France : Larousse.

LAMBERTERIE, I. de (Éd.). (1988). *BRISES. Bulletin de recherches sur l'information en sciences économiques humaines et sociales, ISSN 0293-7166. Appropriation et circulation de l'information*. Paris, France : Institut de l'information scientifique et technique.

LANGLARD, H. (1928). *La liaison dans le français*. Paris : Librairie ancienne Edouard Champion.

LE BORGNE, C., & CORDEREIX, P. (2005). *Réflexion sur la stratégie de constitution et diffusion d'un corpus d'enregistrement sonore extrait des archives de la BPI*. Villeurbanne, Rhône, France.

Le Deuff, O. (2014). *Le temps des humanités digitales: la mutation des sciences humaines et sociales*. (O. Le Deuff, éd.). Limoges, France: Fyp éditions, impr. 2014.

« Le projet Elicop ». (s. d.). Consulté 18 juillet 2015, à l'adresse http://bach.arts.kuleuven.be/elicop/

LORENZ, +O. (1994). « Toménina ». In *Mat y a sé Lila*.: Ga+o-Adri eds, Orléans / Paris, France.

MALÉCOT, A. (1975). « French liaison as a function of grammatical, phonetic and paralinguistic variables ». *Phonetica*, *32*, 161-179.

MALLET, G. (2008). *La liaison en français : descriptions et analyses dans le corpus PFC* (Thèse de doctorat). Université Paris Ouest Nanterre la Défense.

MASTROMONACO, S. M. (1999). *Liaison in French as a second language* (Thèse). University of Toronto, Toronto.

MINEL, J.-L. (2002). *Filtrage sémantique de textes: problèmes, conception et réalisation d'une plate-forme informatique* (HDR). Paris, France.

MINEL, J.-L. (Éd.). (2009). *Filtrage sémantique: de l'annotation à la navigation textuelle*. Paris, France : Hermes science publications.

MINEL, J.-L. (2014). « Les perspectives du Web sémantique pour les SHS au niveau international ». In *Programme ANF - Le Web sémantique pour les SHS*. Fréjus, France. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-01069150

MONDADA, L. (2000). *Décrire la ville: la construction des savoirs urbains dans l'interaction et dans le texte*. Paris, France : Anthropos.

MONDADA, L. (2005). *Chercheurs en interaction : comment émergent les savoirs*. Lausanne, Suisse : Presses polytechniques et universitaires romandes.

MONDADA, L. (Éd.). (2008). *Verbum (Nancy). La pertinence du contexte*. Nancy, France : Presses universitaires de Nancy.

MONDADA, L., & RENAUD, P. (Éd.). (2001). *La linguistique à l'épreuve du terrain urbain*. Bâle, Suisse : Université de Bâle, Romanisches Seminar.

NARDY, A., & BARBU, S. (2006). « Production and judgment in childhood: the case of liaison in french ». In F. Hinskens (éd.), *Language variation - European perspectives. Selected papers from the third international conference on language variation in Europe (ICLaVE3)* (p. 143-152). Amsterdam, Philadelphia : John Benjamins.

NGUYEN, N., WAUQUIER, S., LANCIA, L., & TULLER, B. (2007). « Detection of liaison consonants in speech processing in French: Experimental data and theoretical implications ». In P. Pietro, J. Mascaro, & M. J. Solé (éd.), *Segmental and Prosodic Issues in Romance Phonology* (p. 3-23). Amsterdam : John Benjamins.

NIEL, F., & LABORIER, P. (2009). *Les vicissitudes de l'État linguiste ou Comment les langues minoritaires deviennent l'objet d'une politique sociale linguistique : contribution à une sociologie historique du capital informationnel d'état*. Lille, France: Atelier national de reproduction des thèses.

PIERREL, J.-M. (Éd.). (2005). *TIC et sciences cognitives*. Paris, France : Hermès science publications : Lavoisier.

SALAÜN, J.-M., & HABERT, B.. (2015). *Architecture de l'information : Méthodes, outils, enjeux* (1re éd.). Louvain-la-Neuve; Paris: De Boeck Université.

SCHEER, T. (2004). « Le corpus heuristique : un outil qui montre mais ne démontre pas ». *Corpus*, *3*. Consulté à l'adresse http://corpus.revues.org/210

SPINELLI, E., & MEUNIER, F. (2005). « Le traitement cognitif de la liaison dans la reconnaissance de la parole enchaînée ». *Langages*,*158*, 79-88.

STIVERS, T., MONDADA, L., & STEENSIG, J. (Éd.). (2011). *The morality of knowledge in conversation*. Cambridge, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, Etats-Unis d'Amérique : Cambridge University Press.

STOECKEL, V., & SICCARDI, A. (2004). *Liaison et segmentation des mots chez un enfant québécois de 44 mois - Exploitation d'un corpus dense* (Mémoire de maîtrise). Université Stendhal, Grenoble.

THIBAULT, P., VINCENT, D., & AUDET, G. (1990). *Un Corpus de français parlé: Montréal 84, historique, méthodes et perspectives de recherche*. ERIC Clearinghouse.

TRANEL, B. (1995). « Current issues in French phonology: Liaison and position theories ». In J. A. Goldsmith (éd.), *The Handbook of phonological theory* (p. 798-816). Cambridge: Oxford: Blackwell.

TRANEL, B. (1998). « French liaison and elision revisited: a unified account within optimality theory ». In C. Parodi, C. Quicoli, M. Saltarelli, & M. L. Zubizarreta (éd.), *Aspects of Romance Linguistics - Selected papers from the linguistic symposium on Romance languages (XXIV)* (p. 433-455). Waschington, D.C.: Georgetown University Press.

TUFTE, E. R. (2005). *Visual explanations: images and quantities, evidence and narrative*. Cheshire, Conn., Etats-Unis d'Amérique: Graphics Press.

WANG, H. S., & LIU, H.-C. J. (2010). « The morphologization of liaison consonants in Taiwan Min and Taiwan Hakka ». *Language and Linguistics*, *11*, 1-20.

WAUQUIER-GRAVELINES, S., & BRAUD, V. (2005). « Proto-déterminant et acquisition de la liaison obligatoire en français ». *Langages*, *158*, 53-65.

WAUQUIER-GRAVELINES, S., ENCREVÉ, P., & SCHEER, T. (2005). « Liaison in French, towards an unified explanation of variation ». Colloque *Phonologie du Français Contemporain, Phonological Variation, the case of French*, Tromsø, Norway, 25-27 août 2005.

WAUQUIER, S. (2009). « Acquisition de la liaison en L1 et L2: stratégies phonologiques ou lexicales ». *Aile* , 93-130.

WAUQUIER, S. (s. d.). « Des ornithorynques et des consonnes doublement flottantes. Pour une théorisation unifiée de la liaison ». In *Hommage à Pierre Encrevé*.

Résumé

Observatologie :
Vers une science de l'adéquation observationnelle en linguistique

Ce texte présente les éléments d'un parcours de chercheur vers une science de *l'observation linguistique*. Il s'agit de porter un regard réflexif et de situer la cohérence de différentes activités de recherche qui construisent, in fine, un chemin non linéaire à partir des questions des données de la recherche en linguistique vers la construction et la définition d'un objet scientifique.

Résolument ancré dans le champ d'une linguistique qui refuse d'établir une dichotomie entre la linguistique et la sociolinguistique parce qu'il part du principe que la langue est *par nature sociale*, ce travail est orienté vers la quête de l'adéquation observationnelle. En dépassant le cadre de l'enquête sociolinguistique d'un côté et de la linguistique de corpus de l'autre, il interroge ce que peut être une pratique et une théorie des données linguistiques et de leur condition de production

Le cheminement retracé dans ce document se situe au confluent de l'enquête linguistique et de la linguistique de corpus dans une période épistémologique qui correspond, depuis une dizaine d'années, à l'émergence du domaine des « *humanités numériques* ». Au cœur de cette approche il y a la part essentielle que prennent les données dont on ne peut séparer la collecte de l'exploitation, la méthodologie de la théorie, le terrain de l'analyse, la science de la politique. Leur convergence dessine les contours d'une véritable *science de l'observation* des données linguistiques.

Abstract

Observatology :
Towards a science of observational adequacy in linguistics

This text presents elements of a researcher's pathway towards a science of the linguistic observation. Our aim is to lead a reflection (on) and to situate the coherence of our different research activities that build, in fine, a non-linear way towards the construction and the definition of a scientific object, starting from the issue of research data in linguistics.

Firmly anchored in the field of linguistics that refuses the dichotomy between linguistics and sociolinguistics, and in line with the principle of the social nature of language, this work is oriented towards the quest of an observational consistency. Going beyond the framework of the sociolinguistic survey on one hand and of the corpus linguistics on the other hand, we question what can be a practice and a theory of linguistic data and the condition of their production.

Thus, the pathway presented in this text is intended to be at the confluence of linguistic survey and of corpus linguistics in an epistemological period that gave birth, ten years ago, to the field of "digital humanities". In the heart of this approach remains the essential share of the data in which we can't separate collection and exploitation, methodology and theory, field and analysis. All those elements should be gathered around a real science of the observation of linguistic data.