HAL
open science

CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO

Lotfi Abouda, Olivier Baude

To cite this version:

Lotfi Abouda, Olivier Baude. CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO. Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation, 2006, Albi, France. halshs-01162506

HAL Id: halshs-01162506

https://shs.hal.science/halshs-01162506v1

Submitted on 11 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO[1]

Lotfi ABOUDA – Olivier BAUDE

CORAL – Université d'Orléans

SOMMAIRE

0. Introduction
1. Quelques considérations sur le linguiste et ses corpus
1.1 Données attestées et situées VS masse de données
1.2 Place des corpus oraux
1.3 Corpus disponibles VS corpus fantômes
2. Les corpus ESLO : de la collecte à l'exploitation d'un corpus oral
2.1 ESLO1 un corpus à reconstruire
2.2 ESLO 2 un corpus à anticiper
3. Choix pour la mutualisation et l'interopérabilité d'un grand corpus oral
3.1 Rôle des métadonnées pour l'interopérabilité
3.2 Des données exploitables : le cas de la transcription
3.3 Corpus mutualisé pour des analyses multi-domaines : le test d'eslomelette
4. Conclusion

0. Introduction

Contrairement aux corpus de français écrit il n'existe pas de grand corpus de français oral disponible pour l'ensemble de la communauté scientifique. La présentation du projet de diffusion des corpus ESLO (Enquêtes Socio-Linguistique à Orléans) à l'ensemble des acteurs de la recherche, qu'ils viennent des sciences cognitives ou de l'anthropologie, de la physique (traitement du signal) ou des études de genre (gender studies), de la dictionnairique ou du TAL, est l'occasion d'interroger les raisons complexes d'une telle situation.

Un regard épistémologique, notamment sur la place des données en linguistique, apporte des éléments d'explication qu'il convient de prendre en compte avant de proposer des pistes pour une méthodologie favorable à l'exploitation de grand corpus oraux. Les principaux choix théoriques et techniques opérés lors de l'exploitation scientifique (numérisation, transcription, annotation, diffusion, analyses) du corpus ESLO1, vu comme étape liminaire à ESLO2, répondent à un objectif précis : participer à la réflexion sur l'évolution des modèles et des méthodes de constitution et d'exploitation des corpus oraux destinés à des finalités linguistiques.

1. Quelques considérations sur le linguiste et ses corpus

Sans oser une présentation épistémologique de la place des corpus en linguistique, nous souhaitons présenter quelques considérations sur l'usage des corpus en linguistique qui sont à l'origine du projet de constitution, d'exploitation et de diffusion des corpus ESLO.

[1] Enquête Socio-Linguistique à Orléans.

1.1 Données attestées et situées VS masse de données

La linguistique connaît actuellement un bouleversement méthodologique amorcé il y a plus de 30 ans. Les possibilités offertes par le traitement automatique du langage et notamment les techniques d'exploitation des documents numériques ont permis des développements théoriques fondées sur l'exploitation de corpus, mettant ceux-ci aux centres de la description et de l'analyse linguistique.
Une ambiguïté demeure cependant. En effet, les corpus étaient utilisés bien avant le développement du domaine du TAL. Travailler sur corpus consistait alors à considérer l'objet d'étude comme une collection ordonnée de productions attestées et situées. Cette définition de l'objet impliquait une démarche empirique de description des faits qui s'opposait à une démarche hypothético-déductive fondée sur l'intuition du chercheur. La méthodologie de travail sur corpus était donc un acte scientifique fort et fondateur de certains domaines (sociolinguistique, analyse de la conversation, ethnolinguistique, etc.) centrés sur la conception de "corpus de langue parlée".
Depuis les années 1980, la linguistique de corpus s'est définie autour de grands corpus de langue écrite traités informatiquement comme l'ont décrit Kennedy (Kennedy, 1998) pour l'anglais et Habert pour le français (Habert et al., 1997).

Ainsi les possibilités offertes par le traitement informatique de masse de données sont devenues l'atout principal de la linguistique de corpus. Toutefois trois questions, selon nous centrales, se trouvent biaisées dans ce contexte :

- le corpus est souvent constitué de productions très normés (romans, articles de presses, textes officiels) dont le traitement requiert une standardisation (orthographe, étiquetage, etc.) ;
- le corpus est souvent considéré comme représentatif d'une hétérogénéité des pratiques de part le simple fait qu'il constitue une masse de données ;
- la disponibilité de vaste corpus (FRANTEXT) permet dans de nombreux travaux d'éluder la question pourtant centrale de la constitution du corpus comme première étape d'une théorie linguistique.

1.2 Place des corpus oraux

Si la linguiste du corpus s'est massivement développée, force est de constater que la linguistique dispose de peu de corpus oraux. C'est un paramètre qu'il est facile d'expliquer : la tradition littérale est continue depuis l'Antiquité quand les modes de conservation du son ont moins d'un siècle et demi d'existence. Mais ce n'est pas l'unique raison. L'oral s'accommode beaucoup moins d'un traitement excluant les variations. L'écrit est normalisé par sa présentation même en chaîne de caractères. Il est le produit d'une transcription déjà effectuée, que la source en soit assignée au mental ou au signal. Avant tout retravail par les instruments et les outils du TAL, une homogénéisation de la présentation et des formes a été accomplie à divers niveaux : orthographe, découpe des mots et des phrases, ponctuation…

La recherche en intelligence artificielle a été facilitée, quand elle se donnait les langues pour objet, par la saisie d'énoncés écrits, avec pour conséquence l'élaboration de techniques et d'approches dont l'extension à des corpus oraux (enregistrements, transcriptions phonétiques) était malaisée. On est en présence d'un cas d'école concernant l'ajustement réciproque des données et des outils qui pâtit de l'extension des processus à de nouvelles catégories d'objets.

Les problèmes de l'extension des méthodes éprouvées sur des corpus scripturaux à des corpus oraux se situent sur différents plans :

- insuffisance des corpus oraux, que ce soit en termes quantitatifs de disponibilité globale, ou qualitatifs de fiabilité scientifique ou de prétraitement ;

- dissymétrie des champs d'application de l'enquête (opposition des études linguistiques de terrain - field linguistics), orientées vers les langues sans tradition écrite, et des linguistiques de bureau (armchair linguistics), centrées sur les textes de référence et l'écrit -, les départements informatiques étant plus souvent confrontés à celles-ci pour lesquelles existe de surcroît une forte demande des industries de la langue ;
- parcellisation des enquêtes et des standards retenus pour la collecte, la conservation et la codification : ainsi, le très important travail d'archive orale entrepris dans les deux dernières décennies par les historiens et les sociologues a souvent été entrepris sans finalité externe qui aurait pu assurer une exploitation ouverte des fonds ;
- faible exigence de prescription : les corpus sont constitués sur des objectifs *ad hoc*, ciblant leur finalité en fonction d'objectifs circonscrits, par exemple la reconnaissance vocale ou la fouille de textes ;
- pratiques lacunaires de catalogage et de description des ressources : la bibliothéconomie des archives sonores reste aujourd'hui encore balbutiante et c'est un chantier international où il importe que soient formulées des propositions pour tout ce qui a trait à la standardisation des produits, à l'indexation et à la consultation (représentativité des éléments de catalogage par rapport aux contenus en fonction de pertinences multiples).

1.3 Corpus disponibles *VS* corpus fantômes

Nous l'avons précisé, si la linguistique de corpus s'est considérablement développée, ce n'est pas pour autant, et le fait mérite d'être grassement souligné, que les corpus eux même soient disponibles. En effet, à l'exception notamment de FRANTEXT et du *Monde*, les corpus sont toujours évoqués dans les travaux, mais ne sont que très rarement diffusés. Ils jalonnent les articles et les thèses comme les fantômes hantent les couloirs et les tours : toujours évoqués comme preuve mais n'apparaissant à nul autre qu'à celui qui en parle.
Cette situation ne mérite pas d'être caricaturée et on peut esquisser une typologie des corpus en fonction d'un critère de disponibilité :

- Certains corpus ont été constitués dans le cadre d'une recherche précise et n'ont de pertinence que pour celle-ci. Les conditions de collecte ou le travail très spécifique d'annotation ne permet pas la diffusion de ces données.
- D'autres corpus ne sont pas disponibles par volonté des chercheurs qui souhaitent garder une priorité scientifique sur un travail de collecte couteux et laborieux.
- Enfin il existe des corpus conçus comme des bases de données qui prennent le statut de corpus de référence par le simple fait que ce sont les seuls disponibles. Ces corpus sont alors utilisés simplement … parce qu'ils sont là.

Nous ne pouvons terminer cette courte typologie sans évoquer les corpus totalement fantômes qui fondent certains travaux sans qu'aucune information ne précise les raisons de l'absence de l'accès aux données, pourtant seule garantie d'un travail scientifique en principe ouvert à la falsification.
Le programme ESLO se situe résolument dans une démarche scientifique pour laquelle un corpus non disponible n'existe pas.

En bref, la linguistique de corpus a dans un premier temps peu pris en charge le domaine de la langue parlée et des données situées. Cependant les technologies récentes permettant de numériser le son et d'avoir une synchronisation temporelle entre le signal et une ou des transcriptions ainsi que les initiatives de normalisation de structuration des corpus (*TEI*), des métadonnées (*Dublin core* et *Olac* par exemple) et des données liées ouvrent de nouvelles perspectives pour la linguistique de corpus. Toutefois il n'existe pas

actuellement de grand corpus français de langue parlée disponible pour la communauté scientifique. Le projet ESLO (cf. infra) souhaite répondre à cette demande.

2. Les corpus ESLO : de la collecte à l'exploitation d'un corpus oral

2.1 ESLO1, un corpus à reconstruire

L'Enquête Socio-Linguistique à Orléans (désormais : ESLO1) a été conduite en 1968 par des universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Il s'agit d'un vaste corpus estimé à plus de 300 heures (environ 4 500 000 mots).

Elle comprend environ 200 interviews, toutes référencées (caractérisation sociologique des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien), mais aussi une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.). Certains des enquêtés ont ainsi été enregistré dans des situations très différentes. ESLO 1 couvre l'ensemble des catégories socioprofessionnelles, hommes et femmes, avec plusieurs locuteurs originaires de différentes régions. C'est un échantillon des formats de la communication, des tâches linguistiques, des types de discours selon une approche essentiellement dialogique. Ce corpus représente, par son ampleur, sa rigueur et sa cohérence, le plus important témoignage disponible sur le français parlé avant 1980. Si les fins de sa constitution étaient linguistiques, ESLO 1 est un témoignage unique sur les jugements concernant mai 68 vu de la province ou sur les représentations collectives de la cité à cette époque.

Le Coral (Centre orléanais de recherche en anthropologie et linguistique) a réussi à récupérer en 1993 l'ensemble de documents originaux composés des bandes magnétiques, un catalogue dactylographié, quelques centaines de feuillets de transcription manuscrites (d'une qualité inégale) et les fiches d'identification des locuteurs.

L'opportunité offerte par la numérisation des originaux arrivés en fin de vie a permis au Coral de consacrer un projet à la conservation et à la valorisation du corpus. L'opération de numérisation n'était pas en l'occurrence anodine ; c'est une véritable reconstruction du corpus et sa transformation en un nouvel objet scientifique qui a été opérée. Les documents sonores ont été récolligés et complétés (la conservation avait été défectueuse), numérisés à partir des enregistrements et une indexation et un premier catalogage informatisé a pu être réalisé. Parallèlement, l'exploitation exhaustive d'un sous-ensemble a été entreprise au point de rencontre de données linguistiques variationnistes et cognitives (description d'une tâche). L'étape suivante consiste à transcrire et baliser l'intégralité du corpus.

L'enjeu de cette reconstruction n'est pas neutre. Il s'agit d'établir des principes ayant valeur de normalisation afin de mettre l'ensemble des données à la disposition de la communauté scientifique dans un format qui en permette une exploitation fiable, optimale et intensive, y compris pour des applications industrielles après sélection des contenus.

2.2 ESLO 2, un corpus à anticiper

En partant des acquis d'ESLO 1, une nouvelle enquête, dénommée ESLO 2, a été mise en chantier par le CORAL. Il s'agit, à quarante années de distance, de constituer un corpus comparable dans le produit attendu et dans les modalités de la collecte : l'objectif a été fixé à 400 heures environ de documents sonores qui totaliseraient approximativement 6 000 000 de mots. Réunis, ESLO 1 et ESLO 2 formeront une collection de 700 heures d'enregistrement, soit plus de 10 000 000 de mots, ce qui est considéré aujourd'hui comme une valeur repère pour les investigations projetées.

ESLO 2 a été conçu pour préfigurer la référence attendue dans un domaine qui en est encore à se structurer et dans lequel se manifeste de manière récurrente une demande de définition pour un format standardisé de *collecte*, de *conservation*, de *traitement* et d'*analyse* :

- la *collecte* sur le terrain est première, non seulement dans ses aspects techniques, aujourd'hui bien maîtrisés, mais dans la définition du profil de l'échantillon représentatif et dans la problématisation des interactions entre les témoins et les enquêteurs ;
- la *conservation*, qui inclut la préservation des supports, l'indexation des contenus et l'accessibilité (c'est-à-dire la protection) des données, conditionne le partage des sources à des fins d'étude scientifique ou didactique ;
- le *traitement*, en lien étroit avec le développement des matériels et des langages informatiques, suppose la maîtrise d'une chaîne d'opérations, depuis la conversion numérique des enregistrements jusqu'à une transcription balisée et ouverte à l'ensemble des interrogations pertinentes pour les demandes du linguiste, du sociologue ou des décideurs, des didacticiens voire du grand public ;
- l'*analyse* constitue l'épreuve des théories (et des logiciels) puisqu'elle compare les formalisations et les opérations et qu'elle valide ou infirme les hypothèses en prenant argument de leur compatibilité aux faits.

Les acquis en matière de conservation, de traitement et d'analyse seront reportés sur ESLO1 comme le requiert la comparabilité attendue.

3. Choix pour la mutualisation et l'interopérabilité d'un grand corpus oral

Quels sont les choix et les enjeux contenues dans l'objectif de mutualisation et d'interopérabilité d'un grand corpus oral de type ESLO ? Nous nous bornerons ici à présenter une démarche suffisamment générale qui interroge l'exploitation des corpus en sciences humaines.

3.1 Rôle des métadonnées pour l'interopérabilité

Les corpus constituent des ressources numériques sur lesquelles se fondent la majorité des travaux en linguistique actuellement sans que la question de la forme de ceux-ci et des limites qui bornent cette collection ordonnée soit toujours clairement résolue.

Un corpus est constitué de données brutes et/ou annotés (Véronis 2000, Habert et Fuchs 2004). Dans le cas des corpus oraux, les enregistrements de la parole constituent les données primaires, la transcription et les autres annotations éventuelles représentent des données secondaires. L'ensemble de ces données sont décrites par des métadonnées chargées de documenter le corpus.

Ce sont ces dernières informations, particulièrement importantes pour rendre une ressource disponible mais aussi pour expliciter les critères de sélection et d'organisation des données et donc des bornes du corpus, qui manquent souvent dans les corpus disponibles. Or il s'agit ni plus ni moins de poser ainsi la question de la représentativité du corpus. C'est en effet l'explicitation des bornes du corpus (conditions de productions, de réception, contexte des usages, informations sociologiques sur les producteurs, genre, etc.) qui permet de juger de la représentativité de corpus qui du statut d'échantillon de la langue passe très souvent à celui de corpus de référence (même si ce statut référentiel est implicite) sans aucun regard réflexif sur la forme de celui-ci.

Dans le cas du corpus ESLO1, l'équipe a souhaité conserver le travail de catalogage et de documentation déjà anticipé par les auteurs du corpus. La démarche actuelle et validée par la communauté consiste à utiliser des métadonnées *Dublin Core* et les extensions préconisées par le programme OLAC. Ce jeu d'étiquettes permet des opérations de

catalogage tout à fait satisfaisantes en termes de description d'une ressource qu'on souhaite répertorier pour la rendre accessible.

Cependant cette procédure n'est pas suffisante pour documenter un corpus conçu comme un réservoir qui doit permettre à un chercheur de construire son propre corpus répondant aux exigences de sa recherche. Dans le cadre de cette extraction/construction, les informations permettant de borner le corpus doivent répondre à une granularité très fine. Dans le cas du corpus des ESLO, nous avons déjà pointé l'importance accordée aux informations sur les locuteurs, la situation de collecte et l'échantillonnage dont le but était de constituer un corpus représentatif.
Ce travail méthodologique permet de considérer que la collecte a correspondu à un travail méthodologique rigoureux, source de données représentatives et qu'ainsi le chercheur est sûr d'être confronté à des données pertinentes. Il convient d'aller plus loin et de considérer qu'un corpus doit contenir une riche documentation sur les données mais aussi sur les contextes de production de ces données. Ces contextes concernent aussi bien les données sur les locuteurs et la situation de collecte que l'explicitation de la démarche du chercheur.

Quelles sont les possibilités pour mettre à disposition un corpus qui contient en lui-même les informations sur ses bornes constitutives ? Le standard XML offre des éléments de réponse.
Techniquement ce standard sépare la représentation physique et logique des documents (les données et les métadonnées). Tout document XML comporte donc l'identification des éléments possibles et leurs relations possibles (Définition de Type de Document) et les données identifiées selon cette DTD. C'est alors la notion même de données brutes qui est redéfinie. Ainsi la TEI rend obligatoire la constitution d'un header (en-tête) en début de corpus qui recense les informations sur le contexte de production des données. Cependant le chapitre de la TEI consacré à l'oral est actuellement beaucoup trop succinct pour permettre une véritable normalisation de cette démarche.

Il y a donc un enjeu à considérer les métadonnées comme des éléments de description des données au sens linguistique de celle-ci et non simplement en termes de documentation de ressources. Les métadonnées doivent permettre d'expliciter la démarche du chercheur en proposant une description fine des ses choix théoriques "encapsulés" dans des choix techniques. Les opérations de transcription sont en ce sens un exemple particulièrement éclairant.

3.2 Des données exploitables : le cas de la transcription

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

Depuis quelques années, alors que la manipulation du son numérique devenait très aisée (capacité de stockage, rapidité d'accès, débit suffisant pour une transmission en réseau…), des logiciels permettent la synchronisation du son et de la transcription (*Praat*, *Transcriber*, *Winpitch*, *soundedit*, etc.).
Ces innovations ont des répercussions méthodologiques importantes sur le travail du linguiste. En effet, avec des transcriptions alignées sur le signal sonore, l'oral devient physiquement l'objet d'étude et est systématiquement disponible en même temps que la transcription. Le retour aux données peut alors être systématique, ce qui est de nature à faciliter les procédures de vérification, étape essentielle du travail scientifique, malheureusement souvent rendue impraticable de par l'inaccessibilité des corpus.

Parallèlement, la synchronisation, qui permet l'annotation de segments temporels, offre une base de référence pour de la multi annotation et donc de la multi transcription. On peut

concevoir, pour un même segment, une multitude de transcriptions, opérées dans des cadres théoriques distincts et/ou avec des granularités différentes, dont chacune répond à un besoin scientifique spécifique. Ici, la transcription n'est plus la vérité d'un chercheur (au mieux) ou d'un transcripteur, elle devient cumulative.

Face à l'ampleur de la tâche, les choix pour la transcription d'ESLO1 ont été fondés sur la volonté de mettre à disposition une transcription de l'intégralité du corpus le plus rapidement possible sans que celle-ci n'implique une théorie linguistique très déterminée (même si toute transcription est une formalisation impliquant une théorie).
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. 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 : La segmentation se fait sur une unité intuitive de type "groupe de souffle et/ou unité syntaxique pertinente. Le tour de parole a été défini par les changements de locuteurs uniquement, les pauses indiquées automatiquement par leur durée (précision du centième de seconde).

3.3 Corpus mutualisé pour des analyses multi-domaines : le test d'eslomelette

Le groupe du CORAL qui travaille sur ESLO est composé de chercheurs dont les sensibilités théoriques sont diverses, et les domaines de compétence en linguistique assez variés, allant de l'épistémologie à la syntaxe, en passant par le TAL, la phonologie, la pragmatique, la sociolinguistique, etc.
Cette diversité théorique est vue comme un atout majeur pour ce projet, dont l'objectif premier est la mise à disposition d'un corpus, laquelle ne peut pas être conçue sans multi-opérabilité.
Or, quelle meilleure garantie pour un corpus qui se veut disponible pour la communauté linguistique dans son ensemble que d'être conçu par des chercheurs dont les centres d'intérêts sont assez divers ?
Pour mettre à l'épreuve cette première piste, l'équipe a choisi un échantillon de ce corpus sur lequel elle a décidé d'opérer toutes les étapes du travail linguistique, de l'identification du corpus jusqu'à l'analyse, opérée dans des domaines divers (syntaxe, pragmatique, lexique, phonologie…), en passant par l'annotation, qui comporte elle-même différentes phases (transcription, annotation, métadonnées). L'équipe a donc constitué un sous-corpus composé des 90 réponses à la question "comment-faites-vous une omelette ?". Les couples questions réponses ont été transcrits selon les conventions testées. Ces fichiers de transcription ainsi que l'ensemble des méta données constituent une collection de documents intégrés à une base de donnée XML native. Une interface (xquery) a été réalisée dans le cadre du projet GRICO[2] et du CRDO[3] après un travail conjoint d'informaticiens spécialisés dans la gestion de corpus oraux et les chercheurs en linguistique de l'équipe.

Cette première expérience est intéressante à plus d'un égard. D'abord, elle permet de voir sur un petit échantillon toutes les erreurs (d'annotation, de structuration), qu'il est encore temps d'éviter pour la totalité du corpus. Ensuite, elle précise l'utilité d'un corpus situé.

Pour ne donner ici qu'un exemple, on peut citer une recherche en pragmatique opérée par des membres de l'équipe. L'analyse pragmatique de la question de l'omelette montre qu'à partir de la question zéro, telle qu'elle figure dans le questionnaire – i.e. «

[2] Groupe de Recherche sur l'Interopérabilité des Corpus Oraux. Michel Jacobson (Lacito-CRDO) et Richard Walter (Modyco).
[3] Centre de Ressources pour la Description de l'Oral. http://crdo.vjf.cnrs.fr:8080/exist/crdo/

« Comment est-ce qu'on fait une omelette ? Pourriez-vous m'expliquer comment on fait ? » - les enquêteurs, visiblement gênés par la question, développent toutes sortes de modalisation. Après le relevé systématique des différentes marques de distanciation vis-à-vis de la question, qui se distinguent en fait en deux groupes, à savoir d'une part les « stratégies de justification » (évocation des écarts culturels entre la France et l'Angleterre, contrôle de la qualité du son, etc.) et, d'autre part, les « stratégies d'atténuation » (emploi du conditionnel, l'enchâssement de la question, l'emploi de l'atténuation autonymique, etc.), on peut se poser une série de questions que la nature et la structuration du corpus permettent, et qui auraient été tout simplement impossibles ailleurs. Par exemple, y a-t-il dans ce dégradé de modalisation une variable sociologique ? Autrement dit, l'enquêteur utilise-t-il plus ou moins de modalisation selon le profil de l'enquêté (son âge, son sexe, son niveau sur l'échelle AM) ? Ce type de questions, combien intéressante d'un point de vue linguistique, est tout simplement impossible dans d'autres corpus. Autre interrogation : y a-t-il une variable individuelle ? Autrement dit, les enquêteurs se distinguent-ils les uns des autres vis-à-vis de leur relation avec la question ? Et, d'ailleurs, un enquêteur quelconque utilise-t-il au fil du temps que dure l'enquête (en l'occurrence presque un an) les mêmes stratégies de modalisation ? Toutes ces interrogations, et bien d'autres, auraient été fastidieuses ailleurs : ici, elles sont non seulement possibles, grâce à la fois aux outils du TAL et à la disponibilité des métadonnées, mais en plus utiles : par exemple, les interrogations naïves qui viennent d'être évoquées permettent de poser des questions cruciales, concernant la réflexivité de l'enquête, son statut, son degré de figement et d'interaction, etc. Derrière ces questions, il s'agit ni plus ni moins que de poser la question de la pertinence et de la validité de données non situées.

Cet enjeu n'est pas restreint à la pragmatique et à la sociolinguistique, le travail entrepris par des chercheurs aux objectifs très différents permet de tester les possibilités de réappropriation de contraintes méthodologiques (par exemple, la normalisation recherchée par le chercheur en TAL est-elle compatible avec le linguiste variationniste ?).

Conclusion

Les enjeux inhérents à l'exploitation d'un grand corpus oral ne se résument pas à des choix techniques imposés par les outils du traitement automatique du langage et de la linguistique de corpus. L'exemple des corpus d'ESLO ne mettent en évidence que ce qu'on savait déjà : "on ne peut dissocier l'accumulation des données et la critique de leur constitution".

Cette évidence interroge la linguistique sur la constitution même de son objet mais aussi l'ensemble des sciences sociales sur l'exploitation de la masse de données. La réponse passe nécessairement par la maîtrise de la totalité de la chaîne : de la collecte des données à leurs organisation à des fins d'analyses variées.

Bibliographie

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

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

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

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

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

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

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

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

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

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

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

Habert, B., Fuchs, C., (2004) "Introduction le traitement automatique des langues : des modèles aux ressources", *le français moderne traitement automatique et ressources numérisées pour le français*, pp 1-13.

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

Pierrel, J-M, ed, (2000) *Ingénierie des langues*, Paris, Hermès sciences.

Rastier, F. (2004). « Enjeux épistémologiques de la linguistique de corpus ». *Texto !* [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.

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

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

Véronis, J., (2000), "Annotation automatique de corpus : panorama et état de la technique". In Pierrel J-M (ed), pp 111-130.

Wynne M., 2005, *Developing Linguistic Corpora : a Guide to Good Practice*, AHDS, http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm. Visité le 01 juillet 2006.