Verbum XXX, 2008, 4

INTRODUCTION

Jeanne-Marie DEBAISIEUX
Université Nancy 2 & ATILF
Tiphanie BERTIN
Université Nancy 2 & ATILF
Magali HUSIANYCIA
Université Nancy 2 & ATILF

1. CORPUS ORAUX : ÉTAT DES LIEUX

Le British National Corpus[1], méga corpus de 100 millions de mots dont 10 millions de paroles[2], conçu entre 1991 et 1994 et enrichi depuis de deux sous-corpus, le "BNC Sampler" et le "BNC Baby" a servi de modèle pour la constitution de corpus de référence dans de nombreux pays. The Roudledge Handbook of Corpus Linguistics (2010, 118) recense ainsi en Europe le corpus CORIS/CODIS pour l'italien, Le "Czeech National Corpus" pour la langue tchèque, qui comporte une section orale, le "Prague Spoken Corpus" (PMK), comprenant 800 000 mots[3], le "Corpus Internacional do Português" CINTIL-, (80 millions de mots dont 30% de données orales) le "Corpus del Español" qui comporte une dernière tranche de 20 millions de mots dont un quart sur les données orales) le "Corpus Gesproken Nederland" de 10 millions de mots et d'autres portant sur des langues dites minoritaires tels que le "New corpus of Ireland" (NCI) et le "Scottish Corpus of Texts ans Speech". Dans ce panorama, la France constitue encore une exception, notamment en ce qui concerne les données orales[4] et ne recense qu'une petite dizaine de corpus oraux inégalement accessibles. On peut citer ainsi :

1 http://www.natcorp.ox.ac.uk/corpus/index.xml

2 Ce corpus, contrairement au corpus CANCODE (The Cambridge and Nottingham Corpus of Discourse in English), ne donne pas un accès direct aux enregistrements.

3 L'échantillonnage est décrit dans Corpus Linguistics, And International Handbook (2009, 388).

4 La base Frantext (http://www.atilf.fr) propose 4000 textes écrits, pour 80% d'œuvres littéraires et 20% d'œuvres scientifiques ou techniques.

- le projet P.F.C. http://www.projet-pfc.net (400 locuteurs référencés) présente environ 200 000 mots d'entretiens non guidés.
- La base Clapi http://clapi.univ-lyon2.fr/ qui propose 16 heures de données téléchargeable d'interactions
- Le site CFPP2000 http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII/ qui compte environ 350 000 mots d'entretiens auprès de locuteurs parisiens.

À cette liste, non exhaustive, on peut ajouter le corpus de type archive, CorpAix, constitué par l'équipe du Gars depuis les années 70 sous la direction de C. Blanche-Benveniste, qui comporte environ 2 millions de mots mais n'est pas accessible au grand public, ainsi que le CRFP, constitué autour de Jean Veronis (U. de Provence) dont une partie seulement sera accessible[5].

Il est symptomatique qu'un des plus gros corpus de français parlé ait été constitué en Belgique. Il s'agit de la base Valibel, http://www.uclouvain.be/valibel-corpus.html, dont les données sont consultables soit sous forme de fichier son, soit sous forme de transcription et qui comporte environ 4 millions de mots.

Certes, la situation a évolué durant ces vingt dernières années et l'engouement dont les données orales sont l'objet a fait naître quelques initiatives heureuses. Ainsi le CRDO (Centre de Ressources sur les Données Orales) a été conçu pour recueillir les corpus existants et le Guide des bonnes pratiques (Baude, 2006) a rassemblé un grand nombre de chercheurs autour de la constitution, la transcription, la conservation des données orales ainsi que les problèmes juridiques qu'ils soulèvent. Néanmoins, l'éparpillement des données et leur caractère souvent confidentiel ont un impact négatif sur la recherche dans ce domaine.

Comme le signale Benoit Habert, "on reste un peu [en France] dans les prodromes d'une linguistique sur corpus à part entière" (c.p.). L'absence de corpus de référence limite ainsi les possibilités de l'analyse et en affaiblit les résultats par l'impossibilité que l'on a d'en vérifier les hypothèses.

Sans rentrer dans le détail d'une analyse systémique complexe, on peut attribuer cette situation à l'existence de trois facteurs principaux. Le premier concerne la difficulté pour les chercheurs de voir reconnaître l'activité de constitution de données orales en tant qu'activité scientifique. Alors que les différentes étapes de constitution (recueil, transcription), qui sont coûteuses en temps, constituent un enjeu théorique et méthodologique dont l'impact sur l'analyse a été reconnu par les spécialistes[6], elles sont souvent considérées comme des activités secondaires peu gratifiantes[7]. Le deuxième facteur est également de nature institutionnelle. Elle concerne la difficulté

5 Voir CAPPEAU P., SABIO F., BILGER M. & CHANET C. (2004), pour une présentation du corpus.

6 Voir sur ce point l'ouvrage coordonné par Bilger M. (2008).

7 On peut craindre que cette situation soit aggravée par la pression actuelle à laquelle les chercheurs sont soumis aujourd'hui au travers des évaluations AERES.

qu'il y a à mettre en place des coopérations indispensables entre Sciences du langage et spécialistes du traitement robuste de la parole afin d'aboutir à une linguistique outillée, au sens de Habert (2004), seule à même de faire avancer la recherche sur corpus. Le troisième facteur réside dans le statut que la communauté des linguistes accorde à la langue parlée. Comme le remarque Françoise Gadet[8] à propos de la vitalité de la recherche sur corpus pour le français hors de France, dont le recueil a précédé celui du français hexagonal et qui comporte aujourd'hui plus de données, la communauté des linguistes français semble se caractériser par une attitude "ideology-of-the-standard-oriented".

Or les analyses sur corpus, bien que parcellaires du moins en France, interrogent des aspects fondamentaux de la langue. Il en est ainsi de la corrélation entre lexique et grammaire. Dans le domaine français, une des premières mentions de ce phénomène apparaît dans l'article de Moreau (1986), qui présente une étude pionnière d'analyse statistique sur la négation dans un corpus de 15 heures d'enregistrements de 30 locuteurs invités dans une émission radiophonique. L'auteur signale à propos d'expressions telles que *la raison pour laquelle* ou *la situation dans laquelle* :

> "Ce qui caractérise ces diverses expressions, outre qu'elles se conforment en tous points aux diverses règles linguistiques, c'est d'une part qu'elles sont "intuitivement fréquentes" pour reprendre les termes de Gross (1982) […] D'autre part, que le locuteur ne semble pas y utiliser l'entière liberté que lui permet sa langue quant au choix et à la disposition des unités. On aurait donc affaire à des combinaisons dotées d'un statut régulier du point de vue linguistique, mais d'un statut particulier du point de vue psycholinguistique, puisqu'elles ne semblent pas résulter d'une sélection réelle et d'un assemblage effectif d'unités." (139)

Les contraintes lexicales sur les règles syntaxiques ont été illustrées par Claire Blanche-Benveniste dans de nombreux travaux[9]. Elles sont largement traitées dans la littérature anglo-saxonne, en termes de *chunk, formulaic sequence, prefab*, ou *pattern* et obligent à penser, comme le souligne Sinclair 1991, en termes de corrélations fortes souvent dictées par les genres de textes, *the idiom principle*, plutôt qu'en termes de choix illimité de combinaisons, *the open choice principle*.

Un autre phénomène que les analyses sur corpus ont mis en évidence concerne le rapport entre morphologie et syntaxe. Il serait impossible de signaler ici l'ensemble des travaux parus dans le monde anglo-saxon, notamment à la suite du recueil de Haiman & Thompson (1988). Cette problématique a rencontré ces dernières années un écho dans le domaine français, comme le montre la parution du n° 28 de Faits de langues en 2006 et celle du double volume "La parataxe" édité chez Peter Lang (à paraître en 2010).

Au-delà de faits grammaticaux particuliers, l'analyse sur corpus oraux aboutit également à une interrogation sur la nature même des unités en jeu.

---

8 Je remercie l'auteur qui m'a fourni l'article à paraître.

9 Cf. notamment Blanche-Benveniste (1990, 1997, 2000).

La question est au cœur des approches appelées “macro-syntaxiques”[10] et soulève de nombreuses questions.

Ce bref panorama montre comment la recherche sur corpus, en remettant en cause certains concepts fondamentaux de l’analyse, peut contribuer au renouvellement des études sur la langue et combien il est nécessaire de diffuser toute initiative scientifique contribuant à sensibiliser la communauté à la problématique des données orales. C’est le sens de cette publication.

2. PRÉSENTATION DU VOLUME

Ce numéro spécial de *Verbum* est essentiellement constitué d’articles issus des communications présentées lors de la journée d’étude intitulée : “Corpus Oraux : problèmes méthodologiques de recueil et d’analyse de données”, organisée à l’ATILF-CNRS à Nancy le 27 mars 2009[11] par Tiphanie Bertin et Magali Husianycia. L’objectif était de confronter différentes approches méthodologiques tant au niveau de la constitution des corpus que de leur traitement.

La méthode de constitution d’un corpus d’oral, c’est-à-dire de recueil, de transcription et de préparation à l’analyse, est reconnue aujourd’hui comme devant être sous-tendue par une réflexion méthodologique, compte tenu de son impact sur l’analyse des données. Il existe, à l’heure actuelle, une multitude de modalités de recueil et de traitement des données orales qu’il convient d’évaluer. Nous avons donc souhaité organiser une journée d’étude consacrée à ces questions, afin de faire le point sur l’état des recherches sur ce domaine en France et dans la francophonie[12].

Cette problématique a été soulevée dès les premières études sur le français parlé. Elle est ainsi au centre de l’ouvrage de Claire Blanche-Benveniste et Colette Jeanjean, *Le français parlé. Transcription et édition* paru en 1987. Elle continue encore aujourd’hui de préoccuper à la fois les jeunes chercheurs qui découvrent les problèmes méthodologiques du travail sur l’oral et les chercheurs avertis qui réfléchissent à de nouvelles façons de travailler. Depuis une dizaine d’années, de nombreux ouvrages, colloques et autres rencontres en ont abordé les différents aspects : recueil de données orales et transcriptions, bonnes pratiques, analyses “outillées” de grands corpus, archivage, catalogage, codage, etc. Les possibilités d’outillage ont donné lieu à de nouvelles pratiques de constitution et d’analyse de corpus. Il existe aujourd’hui de nombreux logiciels : d’une part, des logiciels d’aide à la transcription, essentiellement orthographique (*Transcriber*), tantôt couplée à de l’analyse multi-modale (*Transana*), tantôt couplée à une analyse de la voix (*Praat*), et d’autre part, des logiciels d’assistance à l’analyse quantitative et/ou qualitative (*Unitex*).

---

10 Cf. en particulier les travaux de Claire Blanche-Benveniste (1990 et à paraître en 2010) et ceux d’Alain Berrendonner (1990 et 2002).

11 Nous remercions ici tous les contributeurs de cette journée.

12 C’est la raison pour laquelle, il nous a paru important d’intégrer à ce volume l’article de Jean-David Bellonie portant sur le français parlé en Martinique.

Il paraissait donc intéressant d'évaluer l'influence de ces technologies sur la constitution et l'analyse des données orales et les réponses actuelles des chercheurs sur une problématique résumée ici en trois questions :

Comment constituer un corpus de langue orale en fonction des objectifs de recherche ?

Comment et avec quels outils analyser un corpus de données orales ?

Quel est l'effet de l'analyse envisagée sur les modalités de constitution et de transcription ?

Des différentes contributions de ce volume, émergent de façon convergente deux types de réponse. L'une prend en compte la diversité : que ce soit en termes de méthodologie, de types de données ou d'analyse, les approches présentées montrent combien il est illusoire de parler d'oral au singulier et combien il est nécessaire de constituer des données à la fois massives et diversifiées en genre afin de mieux appréhender la répartition des structures grammaticales, la sélection de ces structures en termes d'items lexicaux et plus généralement la correspondance entre les faits grammaticaux et les caractéristiques communicationnelles et situationnelles du texte. Le second point que les contributions mettent en évidence concerne la nécessité d'une démarche rigoureusement explicitée face aux différentes étapes de constitution et d'analyse des données, en rappelant l'effet des choix théoriques et méthodologiques sur les données obtenues et les analyses qui peuvent en découler.

L'article de **Françoise Gadet** pose ainsi un regard critique sur la nature des données généralement collectées et formule un constat que les syntacticiens de l'oral partageront sans doute. La quasi absence des données écologiques, c'est-à-dire non sollicitées par le chercheur, favorise une dérive qui consiste à substituer à la diversité textuelle une pseudo-diversité de type sociolinguistique dont la pertinence pour l'analyse des faits syntaxiques reste à démontrer. S'appuyant sur des travaux menés de façon qualitative, elle montre que l'étude des variations syntaxiques ne pourra être menée de façon systématique sans une réflexion préalable sur la nature des données à constituer. Un point de vue similaire est illustré par l'article de **Mireille Bilger**. Après avoir souligné les difficultés que les corpus oraux posent à l'analyse quantitative (problèmes de transcription des amorces et des structures syntaxiques non marquées), l'auteur montre à partir de l'étude du relatif *lequel* et des prépositions qui lui sont associées comment la prise en compte des types de texte permet d'enrichir l'analyse des faits distributionnels, encore trop souvent traités selon une opposition simpliste entre écrit et oral. Elle dégage ainsi des similitudes inattendues entre les emplois des relatifs dans un corpus de presse écrite et ceux relevés dans un corpus oral de discours politiques. C'est également en ayant recours à un corpus échantillonné en entretiens, narrations, conversations, séquences de classe que **Jean-David Bellonie** aborde les faits de variation dans le français parlé de Martinique. Il peut ainsi relever dans des domaines concernant notamment la subordination, les pronoms, les prépositions, des faits de variations intralinguistiques du français de la Martinique, qui permettent de mesurer le degré de stabili-

sation d'un français régional encore trop peu étudié et plaident pour une intégration de tels usages dans les pratiques d'enseignement du français langue étrangère et maternelle.

C'est sur ce dernier domaine que porte l'article **d'Emmanuelle Canut** et **Martine Vertalier**. Après avoir dressé un panorama historique de l'évolution des études en acquisition, les auteurs montrent comment le recours à un corpus d'interactions spontanées entre enfant et adulte peut constituer, en intégrant l'analyse du langage adressé à l'enfant, une entrée nouvelle pour l'étude des faits linguistiques en lien avec le développement cognitif. La constitution en cours d'une base de données, (la base TCOF, comportant plus de 175h d'enregistrements longitudinaux d'interaction adulte-enfant, ouvre la possibilité notamment de concilier approches qualitatives et approches quantitatives dans un domaine où les ressources en français ont jusqu'alors fait cruellement défaut. C'est également en tant que facteur de transformation que **Veronique Traverso** analyse le recours à une base de données pour les analyses sur l'interaction. D'une activité artisanale aboutissant à un corpus figé et le plus souvent non partagé, la recherche évolue vers la production d'un objet évolutif, ouvert à la collectivité et susceptible d'être enrichi selon la visée des analystes. L'auteur expose ensuite comment, par un aller-retour entre l'analyse longitudinale d'un extrait et l'analyse d'une collection d'extraits, un phénomène tel que la coordination des actions peut être repéré et analysé selon ses diverses modalités de réalisation.

La recherche sur corpus oraux en France souffre d'une pénurie de données qui est régulièrement dénoncée par les chercheurs, même si la situation a évolué durant les dix dernières années. La constitution d'une base de données massive et diversifiée paraît pour l'heure encore peu envisageable. Sur la base de ce constat, l'article de **Christophe Benzitoun** propose une solution provisoire permettant, en exploitant les données aujourd'hui accessibles, d'initier un projet de grammaire sur corpus. L'auteur souligne l'importance, pour mener ce travail à bien, d'une sélection minimale, d'un contrôle des méta données et de la normalisation des corpus susceptibles d'être utilisés. Sur ce point, le problème des conventions de transcriptions se pose de façon cruciale. Cette réflexion est au cœur de l'article de **Paul Cappeau**. Après avoir mis en évidence ce que les erreurs de transcription nous apprennent sur le processus de compréhension de façon générale, l'auteur décrit l'impact de telles erreurs sur l'analyse de la langue parlée. Elles alourdissent d'une part la compréhension de données qui sont, par nature, éloignées de nos habitudes de lecture et font aussi courir le risque de voir se développer une représentation négative de l'oral, dont on sait qu'il constitue encore aujourd'hui un objet sensible. L'auteur plaide ainsi pour une démarche rigoureuse de correction avant édition, seule à même de permettre un accès à des données fiables.

Bien que les analyses sur corpus soient l'objet aujourd'hui d'un engouement général, l'ensemble des contributions rassemblées ici soulignent la nécessité d'un regard critique sur les recherches futures dont les avancées reposent crucialement sur la constitution de données partageables et rigoureusement échantillonnées. Cette avancée ne se fera pas sans une collabo-

ration active entre informaticiens et linguistes. Elle constitue aujourd'hui la condition sine qua non d'une linguistique de corpus capable de dépasser l'effet de mode que connaît la langue parlée et de s'imposer à la communauté comme un nouveau paradigme nécessaire à une meilleure compréhension des faits de langue. C'est le choix qui a été fait dans le projet TCOF dont nous présentons dans ce qui suit les grandes lignes.

3. LE PROJET TRAITEMENT DE CORPUS ORAUX EN FRANÇAIS

Il paraissait impossible de ne pas mentionner dans cette présentation un projet en cours d'élaboration à L'ATILF. Il est en effet à la source, bien que de façon indirecte, de l'organisation de cette journée d'étude. Le projet TCOF est né dans les années 2000 de la volonté de chercheurs nancéiens, travaillant sur l'acquisition du langage chez les jeunes enfants et sur l'analyse du parlé des adultes. Il s'agissait au départ de garantir par numérisation la pérennité de données orales collectées pour des travaux de thèse. Ces données ont par la suite été enrichies grâce à la collecte effectuée par différents groupes d'étudiants de sorte que nous avons peu à peu envisagé de rendre ces données disponibles pour la communauté. Le projet a été rendu possible grâce au soutien logistique et financier de l'ATILF. Nous évoquons ici les aspects concernant la constitution progressive du corpus d'adultes[13] qui se caractérise par une démarche centrée exclusivement sur un échantillonnage de données en termes discursifs et s'est déroulée selon trois grandes étapes[14].

3.1. Première étape : le recueil d'entretiens

Suivant la critériologie éprouvée par l'équipe aixoise, nous nous sommes intéressés tout d'abord à des situations permettant de longues prises de paroles de la part des locuteurs. La variation concernait essentiellement les types de discours sollicités par l'intervieweur : explication techniques, récit de vie, d'incident, de voyage. Ce recueil répondait à la nécessité de disposer de données accessibles pour la recherche et de former, par la pratique, des groupes d'étudiants à la problématique du recueil et de la transcription des données orales. Bien que ces données aient été sollicitées par le chercheur, les conditions de recueil et notamment le fait que les étudiants aient majoritairement choisi d'enregistrer d'autres étudiants leur confèrent des caractéristiques assez différentes de celles des entretiens recueillis par exemple dans le corpus PFC. L'âge et la proximité des participants ont ainsi atténué le caractère formel des échanges et ont contribué à surmonter, nous

13 Cf. L'article d'Emmanuelle Canut et Martine Vertalier dans ce volume pour les corpus d'enfants.

14 Nous n'avons jamais souhaité, et nous n'aurions d'ailleurs pas pu le réaliser, un échantillonnage géographique. Il aurait été possible en revanche de constituer un échantillonnage de type socio-démographique, mais cette idée ne nous a jamais vraiment convaincue et il est réjouissant de constater a posteriori que notre position n'était pas incohérente (Cf. l'article de F. Gadet dans ce volume).

souhaitons le croire, le paradoxe de l'observateur. Les deux extraits suivants, tirés d'une part de l'interview d'un adulte expert et d'autre part, d'une interview entre deux étudiants illustrent cette hypothèse.

L1 et tu pourrais nous résumer la vie d'un journaliste une journée

L2 alors la vie d'un journaliste euh une journée c'est un peu réducteur parce qu'en fait ça commence euh en tout cas à Pont à Mousson ça commence relativement tôt le matin et ça finit en général assez tard le soir donc on grignote un petit peu sur la nuit - euh - - la vie d'un journaliste - alors moi j'arrive le matin euh pas très très tôt sur les coups de dix heures dix heures et demie c'est dû d'abord à une une des spécificités du secteur qui fait que le gros de notre travail ne se ne se déroule pas le matin ça arrive naturellement mais c'est pas le gros de notre travail et puis la deuxième spécificité là elle est complètement étrangère à aux fonctions que j'occupe ici c'est que j'habite à soixante quinze kilomètres de là et donc euh le matin quand je mets un pied par terre je suis à à peu près à une heure de une heure du bureau - donc j'arrive sur les coups de dix heures dix heures et demie et la première des choses après avoir dit bonjour à tout le monde puisque naturellement il faut s'efforcer quand on travaille dans une équipe d'être poli au minimum euh si ce n'est convivial euh c'est de lire le journal de la veille en commençant par le concurrent pour savoir ce qu'il a fait ça me semble être la moindre des choses parce que sans que ce soit une référence absolue le but du jeu c'est pas de couvrir exactement les mêmes choses qu'eux

extrait du corpus Corpus "Journaliste"

L1 mais en général tu travailles que le soir ou + pendant la journée

L2 non d'habitude là je leur ai demandé que je ferme euh le soir

L1 hum ouais

L2 tu sais puisque euh la journée je peux pas j'ai cours alors euh moi je travaille le dimanche

L1 hum

L2 le lundi + et le jeudi

L1 ben ça va s'ils tiennent compte de ton emploi du temps en tant qu'étudiante

L2 ouais ouais ça va il y a pas de problème ils tiennent compte

L1 ouais mais tu fais quoi au Mac Do

L2 alors euh ben moi en fait là je suis en cuisine

L1 hum ouais

L2 tu vois euh j'ai été formée cuisine

L1 hum

L2 et puis après par la suite j'ai été formée caisse mais en fait en caisse je me suis formée toute seule + et en cuisine donc euh et ben ce que je fais euh je fais les sandwichs
L1 hum
L2 donc euh ça dépend tu sais tu as plusieurs postes en cuisine + alors tu as le poste tu as le F.C.N la produc
L1 hum
L2 le toaster la garniture + et le grill
L1 c'est quoi le F.C.N
L2 alors le F.C.N c'est là où tu fais cuire tu sais tous les chaussons les nuggets euh + euh les poulets pour faire les Mc Chicken
L1 hum
L2 et euh les filets
L1 hum ouais
L2 les filet-o-fish tu vois
L1 mais ta formation ça t'a pris combien de temps
L2 la formation ça m'a pris euh deux jours en fait
L1 ah d'accord
L2 ouais
L1 c'est rapide
L2 ouais ça a été rapide + et j'avais une période d'essai par contre de de deux semaines tu vois

Extrait du corpus “MacDo”

Sans rentrer dans le détail et sans tomber dans le travers dénoncé ici même par F. Gadet qui consiste à substituer aux critères de genre des paramètres de type socio-démographique[15], on remarque quelques traits syntacti-co-sémantiques dont la répartition diffère dans les deux extraits. Le premier extrait comporte un nombre important de séries de “subordinations”, relevées comme caractéristiques du genre “explication” depuis les travaux de Claire Blanche-Benveniste. Cette organisation est moins représentée en termes de quantité et de complexité dans le second extrait. Ce dernier comporte en revanche 6 configurations avec détachement à droite ou à gauche, alors que le premier extrait en comporte deux fois moins. Une analyse des deux extraits par le logiciel *Lextutor*[16] montre également une différence en termes de densité lexicale. L’extrait journaliste comporte 278 mots dont 133 types différents. L’extrait Macdo comporte 265 mots dont 113 différents. Dans ce dernier, 7 lexèmes apparaissent plusieurs fois ; *cuisine :* 4 occurrences, *FCN :* 3 occurrences, *filets, formation, rapide* : 2 occurrences[17]. Dans l’extrait “journaliste”, en revanche, seuls deux lexèmes présentent plusieurs occurrences ; *journaliste* : 3 occurrences et *heures* : 6 occurrences.

15 Voir F. Gadet dans ce volume.

16 http://www.lextutor.ca/

17 Ces répétitions sont souvent liées aux constructions à détachement formant des configurations à thème “éclaté”.

Cette brève comparaison ne peut tenir lieu d'analyse, mais semble néanmoins indiquer qu'il est possible au sein du genre interview, et ce sur un même sujet, la présentation d'une occupation professionnelle, d'opérer des distinctions sur la base de faits de variation syntaxique.

De façon paradoxale, cette première tranche de données sera plus longue à mettre à disposition. Elle implique en effet un traitement plus complexe. Il est nécessaire d'harmoniser les conventions de transcription qui ont subi des modifications durant ces dernières années. En outre, les corpus ont été transcrit sous Word et doivent être intégralement alignés. Cette opération devrait pouvoir être automatisée partiellement dans un avenir proche, suite à une collaboration établie avec une équipe du Loria (Laboratoire Lorrain d'Informatique et ses Applications, UMR 7503) qui a abouti à la réalisation d'un aligneur automatique[18] dont l'interface permet un réalignement manuel simultané à l'alignement automatique en temps réel[19].

3.2. Deuxième étape : le recueil de conversations

La deuxième étape du recueil des données a porté sur des corpus de conversations. Elle a été facilitée par les nouveaux outils d'enregistrement et de traitement, enregistreurs numériques de petite taille et logiciel de transcription. Là encore, le fait que le recueil des données ait été confié à des étudiants a permis d'obtenir des données dont le degré de sollicitation est parfois très faible. De nombreux étudiants ont en effet choisi de laisser tourner l'enregistreur dans leur milieu familial ou sur des lieux de rencontre. Le premier exemple présente ainsi une conversation autour du prochain repas de Noël :

L3 ben de toute façon > c'est pas compliqué + tu pars sur euh soit euh ben volaille
L1 euh < **volaille c'est le moins cher**
L3 l'oie > + et < **à la limite boeuf**
L2 c'est vol- > vol- volaille porc boeuf c'est < tu as trois solutions
L3 ouais >
L1 le boeuf < est plus cher
L2 euh **le boeuf** >
L3 ben < ouais mais si tu pars sur du rôti du rôti de
L2 **le boeuf** parce que le boeuf le l- le boeuf c'est comme non >
L3 porc ou du rôti de boeuf des trucs comme ça + ça peut être pas mal < ça
L2 si > tu veux cuire chaud si tu veux recuire chaud **boeuf c'est pas possible**

Extrait du corpus "Repas"

18 JTrans (Cerisara, Mella, Fohr, 2009), téléchargeable sur : http://www.loria.fr/~cerisara/jtrans/index.html

19 Voir E. Canut & V. André (à paraître), pour les aspects techniques de cette collaboration.

Il est courant de juger que les corpus de conversations sont peu propices aux analyses syntaxiques. Néanmoins, les conversations présentent, nous le pensons, un intérêt majeur pour l'étude de certains domaines. Ainsi le premier extrait comporte dans un pan de texte assez réduit (135 mots) deux structures de type averbal *à la limite bœuf* et *le bœuf* et deux structures à détachement sans article *volaille c'est le moins cher* et *bœuf c'est pas possible* qui ne sont pas si fréquentes dans les corpus monologaux collectés jusqu'à présent[20]. L'exemple suivant, de 114 mots, extrait d'un enregistrement dans lequel deux locutrices préparent leur matériel pour une activité de broderie, présente un très fort taux de structures à détachement (8 pour un extrait de 114 mots) lié à l'activité effectuée ; le choix des couleurs de fils.

L2 tu vois **il est > assez clair celui-là**
L1 voilà sept cent soixante-dix-huit + justement
L2 ouais ou celui-là
L1 moi je préfère celui-là il est plus joli + personnellement **celui-là il est trop trop < grisé** quand même
L2 je crois que > **celui-là il est trop mauve**
L1 oui un petit peu
L2 c'est un peu trop lilas
L1 **ça j'aime mieux** < ça **ça c'est chaud**
L2 je pense que **celui-là il serait plus** > < ça fait plus
L1 plus gai plus plus >
L2 **bon celui-là on avait dit c'est saumon**
L1 voilà c'est ça + justement < c'est ça
L2 **celui-là c'est du blanc** >
L1 **alors du blanc c'est bon le mètre il est là** le papier < euh euh

Extrait du corpus "Couture"

Ce résultat contraste avec ceux obtenus par Claire Blanche-Benveniste (1994) qui montre que de telles structures sont peu fréquentes dans les corpus monologaux.

Certes, on pourra nous objecter que la situation enregistrée est très particulière. Il paraît néanmoins intéressant d'avoir accès à des telles données qui permettront sans doute, si elles sont assez nombreuses, de réfléchir sur le lien entre activité effectuée et apparition des structures linguistiques, dans la perspective de construction d'une grammaire des usages.

3.3. Troisième étape : le recueil de données écologiques

Cette étape a démarré en 2009 et nous avons été surpris par la capacité de jeunes étudiant(e)s à collecter des données souvent peu accessibles telles que des réunions de travail ou à affronter avec succès des transcriptions que

20 C'est ainsi qu'une étude comparative sur les énoncés verbaux et averbaux dans le corpus italien et le corpus français de CoralRom montre que leur proportion est nettement plus grande dans les conversations que dans les monologues.

nous pensions irréalisables, par exemple de réunion de conseils municipaux ou d'association qui peuvent comporter plus de 10 personnes. Nous possédons même quelques raretés : consultation médicale, entretien avec un juge, plaidoirie, que nous espérons voir se multiplier. Là encore, il sera nécessaire de mener des études détaillées pour appréhender les types de variations susceptibles d'apparaître au sein des sous-genres de cette tranche. Certains faits bien sûr sont attendus, tels que ceux notés en gras dans l'exemple suivant extrait d'une réunion de conseil municipal.

> euh donc **effectivement** euh **par rapport à** cette aire de service de camping-cars il nous reste à à créer une euh une régie + puisque **concernant** la co- la consommation d'eau et d'électricité **notamment** la recharge d'une batterie il nous fallait euh envisager la possibilité de mettre en place un système + nous avons opté pour un système de jetons qui seront disponibles en mairie à l'office de tourisme + alors il il restait à définir le prix + euh là aussi **les services de la ville** ont ont mené un certain nombre d'enquêtes et on a décidé à l'unanimité au niveau de la commission de s'aligner sur les tarifs pratiqués à ¤T1¤[21] + euh ben **pour que** ça soit relativement homogène et **puis qu'on** évite peut-être de faire venir des gens + si on avait pratiqué des tarifs plus bas à ¤T2¤ + parce qu'une fois que ça se sait les gens du ¤T3¤ du ¤T3¤ risquaient **de venir** euh en nombre ici + et là pour le coup **de saturer** notre aire de service donc voilà + à l'unanimité de la commission **nous avons décidé** de fixer ce tarif à quatre euros par jeton + et ça donnera donc lieu à la création d'une régie municipale + **y a-t-il** des questions
>
> Extrait du corpus "Mairie"

On note ainsi l'usage de *par rapport* utilisé pour introduire le thème de l'intervention et que nous retrouvons dans plusieurs enregistrements, l'apparition du *nous* sujet, la forte proportion de sujets lexicaux, la présence de participe présent et d'adverbes en "ment" et de façon générale, la structure normative des constructions, (coordination de subordonnées ou d'infinitifs). Mais il reste encore bien des points à observer afin de déterminer les caractéristiques des sous-genres de parole publique ou professionnelle.

Le projet TCOF, qui occupe toute une équipe[22] depuis maintenant 5 ans a permis de constituer une base de données, qui certes, ne répond pas aux principes reconnus aujourd'hui pour la constitution d'un corpus de référence et s'apparente à une collection de textes, mais dont les modalités de consultation ont été pensées pour permettre aux utilisateurs de construire des corpus de travail rigoureusement paramétrés. La plate forme, commune aux corpus enfants et adultes est hébergée sur le site du CNRTL (ATILF),

21 Toponyme anonymisé.

22 Elle est composée aujourd'hui de Virginie André, Christophe Benzitoun, Emmanuelle Canut et Jeanne-Marie Debaisieux (Université Nancy 2), de Bertrand Gaiffe, Evelyne Jacquey et Etienne Petitjean (CNRS ATILF).

http://www.cnrtl.fr/corpus/tcof/. Elle a été réalisée en collaboration avec l'équipe "Ressources et normalisation" et donne accès à une fiche de méta données dont certains champs sont interrogeables. L'utilisateur aura ainsi la possibilité de télécharger des corpus adultes selon plusieurs paramètres. Certains concernent les locuteurs ; âge et sexe et le nombre de participants, d'autres concernent le document. Ainsi la rubrique *Cadre situationnel* permet de choisir un corpus se rapportant à une situation privée, publique ou professionnelle[23]. La rubrique *Genre de discours* permet ensuite de choisir parmi une liste qui sera progressivement enrichie et qui comporte aujourd'hui 7 étiquettes : *entretien, conversation, réunion, relation de service, débat, cours, conférence, discours politique*.

Les autres informations concernant soit l'enregistrement : date, longueur, nombre de mots, résumé, soit les locuteurs : études, profession, sont consultables mais ne peuvent donner lieu à interrogation. Le site héberge également un concordancier en cours de finalisation au Loria, *JConc*[24], qui permettra d'interroger l'alignement texte et son, à l'instar du logiciel Contexte[25] dont il s'inspire.

3.4. État actuel de la base

Au moment de la rédaction de cette présentation, la base, qui devrait être accessible dans les deux mois à venir, comporte 162 000 mots répartis de façon équilibrée entre les trois rubriques, privée, publique, professionnelle[26] et téléchargeables sur le site. 200 000 mots sont en cours de révision et plus de 100 heures sont en cours d'alignement. La faible proportion des données accessibles aujourd'hui est liée à la volonté de l'équipe de faire en sorte que les corpus soient corrigés par un expert puis anonymisés. La démarche est coûteuse en temps, et ce bien que nous puissions compter sur la collaboration de personnels ITA du laboratoire[27] et d'étudiants vacataires[28] que nous avons formés à la transcription. Comme le souligne Paul Cappeau dans ce volume, cette minutie est pour nous la garantie de données fiables. Bien que modeste, cette mise à disposition nous a paru indispensable. Elle

23 L'étiquette réfère bien à la situation et non au type de parole. En effet on peut considérer qu'une réunion publique n'est pas toujours de nature professionnelle. A l'inverse, un enregistrement effectué à la scolarité de l'université entre un étudiant et une personne de l'administration sera considéré comme professionnel. Les entretiens portant sur des activités professionnelles, dont on sait qu'ils comportent des traits spécifiques, ont été classés sous la rubrique privée, l'utilisateur ayant accès aux contenus abordés au travers du résumé.

24 Voir E. Canut & V. André (à paraître), pour les caractéristiques techniques du logiciel.

25 Réalisé par J. Veronis (U. de Provence).

26 Pour une description des aspects techniques, transcriptions, téléchargement des données, voir E. Canut & V. André (à paraître).

27 Il s'agit de Gisèle Cagne, Isabelle Clément, Josette Frecher, Christiane Jadelot et Françoise Weiss.

28 Stéphanie Houin, Youma Sow et Cécile Desse.

constitue un complément non négligeable aux données aujourd'hui accessibles et nous espérons qu'elle suscite d'autres initiatives. C'est à partir de telles initiatives d'équipe, qu'en l'absence de volonté politique ou institutionnelle au plus haut niveau, il sera possible d'aboutir à la constitution d'un véritable corpus de langue parlée en français.

BIBLIOGRAPHIE

BAUDE O. (coordonné par) (2006), *Corpus oraux, guide des bonnes pratiques*, Orléans, Presses universitaires d'Orléans / CNRS Editions.

BERRENDONNER A. (1990), "Pour une macro-syntaxe", *Travaux de linguistique*, 21, 25-31.

BERRENDONNER A. (2002), "Les deux syntaxes", *Verbum*, XXIV, n° 1-2, 23-36.

BILGER M. (coordonné par) (2008), *Données orales : les enjeux de la transcription*, Cahiers de l'Université de Perpignan, 37.

BLANCHE-BENVENISTE C. & al. (1990), *Le français parlé, études grammaticales*, Paris, Edition du CNRS, coll. Sciences du langage.

BLANCHE-BENVENISTE C. (1994), "Quelques caractéristiques grammaticales des 'sujets' employés dans français parlé des conversations", in Yaguello M. (ed.), *Subjectod and Subjectivity. The status of the subject in linguistic theory*, 77-107, Paris, Ophrys ; London, Institut français du Royaume Uni.

BLANCHE-BENVENISTE C. (1997), "De l'utilité du corpus linguistique", *Revue Française de Linguistique Appliquée, Dossier Corpus. De leur constitution à leur exploitation*, vol. I fasc. 2, 25-42.

BLANCHE-BENVENISTE C. (2000), "Convergences de matériel grammatical permettant d'établir des typologies textuelles", in Bilger M. (coord), *Linguistique sur corpus. Etudes et réflexions*, Cahiers de l'université de Perpignan, 31, 103-116.

BLANCHE-BENVENISTE C. (à paraître en 2010), *Le français : usages de la langue Parlée,* Louvain, Peeters.

BLANCHE-BENVENISTE C. & JEANJEAN C. (1986), *Le français parlé, transcription et édition*, Paris, INALF, Didier édition.

BEGUELIN M.-J., AVANZI M. & CORMINBOEUF G. (éds) (à paraître en 2010), *La Parataxe*, 2 vol, Berne, Peter Lang.

CAPPEAU P., SABIO F., BILGER M. & CHANET C. (2004), *Autour du Corpus de référence du français parlé*, *Recherches sur le Français Parlé*, 18, PUP Aix-en-Provence.

CANUT E. & ANDRE V. (à paraître), "Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français)", in Schmale G. (coord.), *Pratiques* 147-148 : Interactions et Corpus Oraux

FAITS DE LANGUES, n°28, 2006, *Coordination et subordination : typologie et modélisation*, Ophrys

GADET F. (à paraître), "What can be learned about the grammar of French from corpora of French spoken outside France", Actes du 3e colloque "Grammar and Corpora", Mannheim, septembre 2009.

HABERT B. (2004), “Outiller la linguistique : de l’emprunt de techniques aux rencontres de savoirs”, *Revue Française de Linguistique Appliquée*, vol. IX, n° 1, 5-24.

HAIMAN J. & THOMPSON S.A. (eds) (1988), *Clause Combining in Grammar and Discourse*, Amsterdam/Philadelphia, John Benjamins.

LÜDELING A. & KYTO M. (eds) (à paraître en 2009), *Corpus Linguistics. An International Handbook*, (2 vol.), Berlin/New York, Walter de Gruyter.

MOREAU M.-L. (1986), “les séquences préformées : entre les combinaisons libres et les idiomatismes. Le cas de la négation avec ou sans *ne*”, *Le français moderne*, 54, 137-160.

O’ KEEFFE A. & McCARTY M. (à paraître en 2010), *The Roudledge Handbook of corpus Linguistic*s, London/New York, Routledge.

SINCLAIR J. (1991), *Corpus, Concordance, Collocation,* Oxford, Oxford University Press.