CORPUS

Corpus

10 | 2011
Varia

Constitution et exploitation d'un corpus de français parlé parisien

Sonia Branca-Rosoff, Serge Fleury, Florence Lefeuvre et Matthew Pires

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2033
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 novembre 2011
Pagination : 81-98
ISSN : 1638-9808

Référence électronique
Sonia Branca-Rosoff, Serge Fleury, Florence Lefeuvre et Matthew Pires, « Constitution et exploitation d'un corpus
de français parlé parisien », *Corpus* [En ligne], 10 | 2011, mis en ligne le 13 juin 2012, consulté le 04 mai 2019. URL : http://journals.openedition.org/corpus/2033

Constitution et exploitation d'un corpus de français parlé parisien

Sonia BRANCA-ROSOFF*, Serge FLEURY*
Florence LEFEUVRE*, Matthew PIRES**
* Fédération Clesthia-SYLED (Paris 3)
** EA2281 Laseldi, (Univ de Franche-Comté)

Le Corpus du Français Parlé Parisien des années 2000 (désormais CFPP2000) est une base de données sur le français parlé par des adultes de la région Ile-de-France. Le corpus, en cours de construction, et qui devrait atteindre un million de mots, a été constitué à partir d'entretiens sur la relation des habitants à leur quartier. En 2011, dans sa version en ligne, il comprend environ 500 000 mots transcrits orthographiquement et alignés au tour de parole.

Comme le rappelait fortement S. Mellet dans le premier numéro de la revue *Corpus*, en s'insurgeant contre les oppositions trop simples entre une linguistique des « données » attestées et une linguistique d'exemples forgés, la linguistique de corpus est toujours une construction qui choisit et organise ses matériaux. C'est bien en fonction d'un projet linguistique que nous[1] avons conçu CFPP2000. Constatant qu'il n'y avait pas encore assez de ressources disponibles pour développer des analyses sociolinguistiques du français parlé hexagonal, nous avons eu comme premier objectif de combler ce manque. En effet, le Corpaix avait comme visée première un échantillonnage en genres de discours ; le corpus ESLO est un précieux corpus variationniste qui présente une dimension longitudinale, mais Orléans est une petite ville (voir Baude *et al.* ici même). Or, des études portant sur des métropoles comme Londres (voir

1 S. Branca-Rosoff, S. Fleury, F. Lefeuvre (Clesthia-Syled, Paris 3) et M. Pires (EA2281 Laseldi, Université de Franche Comté).

Kerswill & Cheshire, s. d. ; Lodge 2004) ont montré que les variétés langagières pratiquées dans les conurbations sont les moteurs du changement linguistique. Nous avions donc deux objectifs essentiels : rassembler un corpus suffisant pour étudier les choix linguistiques et discursifs opérés par les locuteurs de Paris et de sa banlieue en fonction de leurs caractéristiques sociales ; collaborer avec d'autres laboratoires pour mener des comparaisons à l'échelle de la francophonie sur les dynamiques linguistiques en cours[2].

La conception de CFPP2000 repose sur plusieurs choix que nous allons commenter en abordant d'abord les problèmes liés à l'échantillonnage des locuteurs et au genre de l'entretien. Nous rendrons ensuite compte de « l'outillage » actuel du corpus et de quelques exploitations en cours.

1. Le terrain

1.1 Un corpus construit en vue d'une exploitation sociolinguistique

Le corpus a été construit en vue d'une exploitation sociolinguistique en tenant compte de cinq paramètres, le sexe, l'âge, le diplôme, l'itinéraire professionnel des locuteurs, enfin le lieu d'habitation. Nous savons que ces paramètres sociologiques interagissent avec la construction individuelle des identités et l'évènement communicatif que constitue l'entretien. Il est possible de travailler aussi sur ces dimensions parce que les entretiens contiennent une grande partie des ressources nécessaires pour les nuancer ou les corriger (Fleury & Branca 2009). La complexité des identifications des locuteurs conduit par ailleurs à ne pas mythifier l'équilibrage de l'échantillon. Lorsqu'on raisonne sur une centaine de personnes, il serait risqué de décider qu'une poignée de locuteurs est représentative de sa catégorie sociale[3]. L'important est de maintenir une exigence de diversification des locuteurs interrogés.

2 Voir déjà Barbéris 2010 pour une comparaison du *tu* générique dans CFPP2000 avec les résultats obtenus par Sankoff et Laberge sur le corpus de Montréal.

3 Enregistrer un groupe de trentenaires qui se réunit depuis des années pour jouer au poker, c'est rendre saillante cette « identité », alors même que

Si on compare CFPP2000 avec des corpus anglais récents, on remarque d'autres différences importantes. Le Bergen Corpus of London Teenage Language (COLT)[4] est constitué d'auto-enregistrements de conversations entre les enquêtés et leurs pairs, qui permettent de se faire une idée du vernaculaire des adolescents. P. Kerswill et J. Cheshire, qui s'intéressaient à l'impact du contact entre l'anglais cockney et les variétés des immigrés dans l'émergence de formes syntaxiques innovatrices comme par exemple l'indéfini « man » (= « on » en français), ont mené des entretiens avec des adolescents dans deux quartiers à profil socio-économique opposé, cette polarisation leur permettant de limiter les champs de variation et de faire émerger une opposition entre les vernaculaires populaires et le maintien relatif du standard chez les jeunes appartenant à des communautés plus aisées[5].

Les objectifs de CFPP2000 ont entraîné d'autres choix : le corpus est utilisé en premier lieu pour étudier les modifications qui interviennent dans ce qu'on peut considérer comme un parisien véhiculaire en tension entre le pôle du standard et le pôle du vernaculaire. Il est construit en second lieu pour étudier la situation complexe de Paris où par exemple les « bobos » du Marais représentent une figure des couches privilégiées distincte de celle qu'incarnent les bourgeois du 7^{e} ou du 16^{e} arrondissement[6]. Les couples oppositifs construits par les auteurs de « Linguistic Innovators » auraient écrasé cette

---

dans d'autres contextes, ils apparaîtraient plutôt dans des rôles de musiciens, agents immobiliers, pères de famille, etc. (voir [03-01] à paraître dans CFPP2000).

4 Ce corpus est d'une taille similaire à CFPP2000 – 500 000 mots pour 31 locuteurs. Comme nous, ses concepteurs ont restreint le recueil des données à la capitale parce que les nouvelles tendances linguistiques émergent dans les capitales : « from where it can be expected to spread to the rest of the country, and even abroad » (*User's Manual*, p. 1).

5 Le corpus se nomme *Linguistic innovators* ; voir la page : http://www.lancs.ac.uk/fss/projects/linguistics/innovators/index.htm.

6 Les exemples tirés de CFPP2000 sont identifiés par le quartier concerné et le numéro de l'entretien (identifiant défini dans CFPP2000) ; dans [11-03], 11 renvoie à l'arrondissement (le 11^{e}), 03 au troisième enregistrement réalisé.

diversité linguistique d'autant plus intéressante que les mots à la mode et la prononciation un peu relâchée des Parisiens branchés ont peut-être plus d'influence sur les nouvelles générations que la version la plus traditionnelle du standard.

1.2 Une première tranche privilégiant les locuteurs natifs

Une contrainte forte a été introduite afin de mesurer l'influence éventuelle sur les parlures des enquêtés du lieu d'habitation et de scolarisation à l'adolescence. L'un des enquêtés au moins devait être né dans le lieu d'enquête ou y être arrivé avant l'adolescence. Il y a là évidemment une forte idéalisation des données et il nous faut rendre compte de « l'effacement symbolique de tout ce qu'il [le corpus] ne contient pas[7] ». En particulier, les témoins « principaux » nés hors de la région parisienne ont été écartés de la première tranche du corpus alors que les Parisiens natifs représentent moins de 25 % des 2 125 246 habitants de Paris recensés en 1999[8] et alors que Paris – comme toutes les grandes villes du monde – est une ville plurilingue dans laquelle les mouvements migratoires ont généré des situations de contact du français avec les langues d'origine des locuteurs. Seulement, on rapporte parfois un peu rapidement la naissance des innovations linguistiques à la multiethnicité. En portant d'abord l'attention sur les Parisiens d'origine, nous voulions nous donner les moyens d'évaluer la variabilité ordinaire de leurs parlers, qu'elle soit de nature phonétique, morphosyntaxique ou lexicale. Cet objet en un sens abstrait, obtenu en suspendant l'examen des conséquences des échanges interlangues, doit permettre en un second temps de mieux situer les ruptures éventuelles dues au multilinguisme de locuteurs dont le français est la langue seconde, et ce qui est rapportable à la variation linguistique présente chez les locuteurs français monolingues.

7 Ce que R. Nicolaï a appelé le « paradoxe de l'archéologue » (cité dans Mellet 2002).

8 Paris intra-muros représente 3,7 % de la population de la France. Le critère de l'enfance à Paris n'a été respecté que pour le locuteur choisi comme témoin principal. Une certaine diversité est réintroduite lorsque les locuteurs 2, 3, etc. viennent d'ailleurs. Tel est aussi le cas des enquêteurs qu'ils soient anglais (M. Pires) ou provinciaux (F. Lefeuvre, S. Branca-Rosoff).

C'est pourquoi est d'ores et déjà amorcée une seconde tranche d'enregistrements. Le corpus principal sera complété par quelques études de cas permettant de mieux situer les variantes relevant d'une simplification d'un français resté une langue étrangère.

2. L'entretien

La rançon d'une orientation sociolinguistique est qu'il faut – au moins dans un premier temps – normaliser les données pour les rendre comparables.

2.1 Un genre, l'interview semi-préparée

Les moyens n'étant pas extensibles, l'enquête a été limitée à un genre spécifique que nous appelons indifféremment, comme dans l'usage ordinaire, interview, entretien ou entrevue. Le guide d'entretien dont nous sommes partis a permis de demander aux locuteurs d'accomplir des tâches discursives relativement comparables dans des situations à peu près équivalentes. Pour les mêmes raisons, nous nous sommes centrés sur un thème, la relation des personnes interrogées à leur quartier et plus largement à Paris ou à leur banlieue. Nous n'avons donc pas l'ambition de fournir « un corpus de référence » du français parlé en général (Sinclair 1991 ; Bilger éd. 2000). Comme tous les genres, la situation d'interview est caractérisée sur le plan externe par sa dimension sociale et communicationnelle. Elle instaure une distribution des rôles bien intériorisée dans la France contemporaine entre l'enquêteur qui pose des questions et l'enquêté qui y répond (Branca-Rosoff 1999). Par ailleurs, l'entretien réflexif incite à de longues prises de parole et à une riche activité explicative et argumentative propice au développement d'une syntaxe complexe.

2.2 Des contrastes entre dialogues et multilogues ; enquêteurs à ethos empathique et enquêteurs plus distants

Le corpus a été diversifié en fonction du nombre de locuteurs, tantôt réduit à deux, tantôt élargi à plusieurs : les multilogues favorisent l'expression de l'accord et des désaccords entre participants et de façon générale l'argumentation spontanée (ex. [11-03]), alors que le dialogue d'un enquêté et d'un enquêteur

aboutit à des corpus qui conviennent bien à l'analyse prosodique et qui comportent davantage de longues séquences où l'enquêté parle seul, encouragé par les « mm mm » approbatifs de son partenaire (ex. [07-05]).

Par ailleurs, les enquêtes sont interprétées comme des *co-constructions* qui associent l'activité langagière de locuteurs ordinaires et l'activité questionnante des chercheurs, ce qui invite à tenir compte du style propre de chaque participant. Comme l'ont fortement souligné les ethnométhodologues, le positionnement de l'enquêteur peut entraîner des variations dans la représentation de la rencontre en train de se dérouler. Il est ainsi intéressant d'observer les effets du contraste entre l'ethos de F. Lefeuvre, assez respectueuse du cadre du questionnaire, et l'ethos relativement familier de S. Branca-Rosoff, soucieuse de manifester sa sympathie aux personnes interrogées, entrant dans la dynamique de l'échange, quitte à déclencher de longues digressions. Il est aussi frappant de la voir abandonner la neutralité de l'enquêteur pour suggérer de temps en temps ses opinions, alors que les interventions de M. Pires sont plus conformes à l'idée classique que l'on se fait d'un enquêteur qui ne doit pas influencer ses co-locuteurs.

2.3 Des activités linguistiques diversifiées

Un même genre peut comporter une diversité de types de discours. Les locuteurs de CFPP2000 ont ainsi été invités à décrire leur quartier, leurs itinéraires, à donner leur opinion argumentée sur des questions plus ou moins polémiques, à raconter des anecdotes, à s'exprimer de façon métalinguistique sur les langues, particulièrement sur le français, à s'exprimer à propos de leur travail[9], ce qui ouvre sur une observation systématique de ces usages du français. Voici un bref exemple de narration :

9 Il y a de nombreux malentendus sur la frontière entre genres et types de discours. Le groupe Langage & Travail, l'UMR ICAR, l'ANR CIEL ont entrepris de collecter des corpus en vue de réfléchir aux relations unissant activité de travail et activité de langage. CFPP2000 ne contient pas de propos échangés lors de situations professionnelles. En revanche, comme c'est par exemple le cas pour de nombreux corpus de CORPAIX, certains locuteurs s'expriment *à propos* de leur activité professionnelle : dans ces passages, apparaissent des formes syntaxiques, lexicales et énonciatives qu'ils manient d'habitude dans

> une anecdote un un jeune j'sais pas si tu t'rappelles + il a craché à la vitre de la porte de l'éco- de de ma loge + je l'ai attrapé je lui ai fait nettoyer + le père le lendemain il est venu me voir + en disant pourquoi j'ai fait nettoyer son fils + + [12-03, Valentine Testanier, femme, 60 ans, concierge]

L'analyse de telles séquences a déjà été bien amorcée par J.-M. Adam (1990). On remarque ici le terme métadiscursif *anecdote* qui fonctionne comme marqueur d'ouverture, les étapes du récit marquées par des constructions verbales au passé avec des sujets humains, le discours rapporté qui sert de marqueur de clôture. Cependant les récits oraux font apparaître régulièrement des sortes de parenthèses explicatives et la chute – qui apparaît dans l'exemple suivant sous la forme d'une phrase évaluative – peut être énoncée par le colocuteur soulignant la nature dialogale de l'activité narrative :

> Yvette Audin : donc quand mes parents ont cherché à + un appartement ils habitaient Paris tous les deux + ils avaient été élevés à Paris tous les deux + ils se sont mariés en + mille neuf cent vingt neuf et + ils ont cherché à se loger dans Paris + *à l'époque c'était très difficile* + on leur a signalé qu'il y avait un appartement à louer éventuellement rue Barbet de Jouy ma mère est venue l'visiter et la propriétaire de l'époque lui a dit euh + « nous pouvons vous le louer à une condition c'est que vous nous promettiez que vous n'ayez pas d'enfants » *cet appartement qui je précise faisait deux cent mètres carrés euh était un appartement considéré pour un couple c'est à dire qu'il y avait euh trois pièces de réception un grand salon un petit salon et une salle à manger + il y avait un grand couloir avec la cuisine au bout et à côté une pièce qui servait de lingerie + et + y avait une chambre pour les + le couple qui habitait et à côté il y avait une autre chambre qui servait de dressing* donc elle

l'exercice de leur profession (voir Blanche Benveniste 2000).

> lui a dit « vous comprenez c'est un appartement qui est beaucoup trop petit pour avoir un enfant »
>
> Enq : mm c'est très joli [07-05, Yvette Audin, 70 ans, a travaillé dans l'édition]

2.4 Le choix d'interviews longues : la variation intra-individuelle

Les enregistrements durent environ une heure parce que la longueur favorise la variation à l'intérieur même de la situation d'entretien : des locuteurs qui emploient des variantes hautes au début de l'entretien baissent la garde et adoptent un ton plus détendu à la fin. Nous observons alors la tension entre un registre haut et un registre vernaculaire, qui correspond à la nature double du français ordinaire à la fois familier et soumis à des normes liées à l'histoire de sa standardisation, même si, chez les locuteurs les plus soucieux de norme, le registre quotidien est quasiment effacé, alors que d'autres mettent en avant cette partie de leur répertoire (on peut contraster ainsi le style tendu de Paul et Pierre-Marie Simo [20-01] et le style relâché de Katia Teixera [11-04].

Des entretiens longs augmentent les chances de rencontrer plusieurs attestations d'une même tournure et font apparaître l'importance des variations idiolectales : par exemple, les 11 occurrences de *en revanche* sur l'ensemble du corpus se rencontrent seulement chez une enquêtée et une des enquêtrices, alors que les autres locuteurs emploient *par contre* (Branca-Rosoff *et al.* 2009).

Enfin, le choix d'entretiens longs permet un va-et-vient entre le niveau macro et le niveau micro de l'analyse. Au niveau macro, il ne s'agit pas évidemment de prédire ce qu'un individu va faire à un moment donné puisque chaque personne a à sa disposition des ressources variées dont elle use d'une façon qui reste largement imprévisible, mais nous pouvons faire des généralisations en termes de probabilité et voir si des traits linguistiques sont corrélables à des propriétés socio-linguistiques. Cette approche passe par une délinéarisation : les interviews sont concaténées et traitées comme un ensemble dont on peut extraire les énoncés qui présentent les traits qui nous intéressent. Au niveau micro de l'interaction, il est possible d'envisager ce qui

est actualisé, éventuellement revendiqué, ou, à l'inverse, évité, refusé. Le texte comme totalité est alors l'unité pertinente approchable en tenant compte de la dynamique de l'échange.

3. Le matériau mis en ligne

Les outils informatiques modifient en profondeur le travail du chercheur.

3.1 Un corpus outillé

La numérisation permet de consulter le corpus à partir des ordinateurs des usagers sous une forme alignant, au tour de parole, son et transcription : l'évolution des outils informatisés a changé, nous semble-t-il, les enjeux de la transcription et nous a confortés dans le choix d'une transcription orthographique dont la base est morphologique[10], proche de celle du GARS ou du corpus VALIBEL. Cette option permet de transcrire rapidement ; elle offre un confort de lecture aux chercheurs, plus grand que les transcriptions qui cherchent à coder des propriétés non segmentales, ce qui est important pour des textes longs ; les concordanciers sont directement utilisables. Les informations orales ne sont pas perdues puisque le chercheur intéressé par des phénomènes phonétiques ou prosodiques peut facilement retourner aux documents sonores.

Les ressources du corpus CFPP2000 sont présentes en ligne sous différentes facettes. Les fichiers de transcription et les fichiers audio associés sont directement disponibles au téléchargement : les fichiers audio sont disponibles aux formats wave[11] et mp3, les fichiers de transcription sont au format Transcriber[12]

10 La seule adaptation de l'orthographe que nous nous sommes autorisée concerne les clitiques. Parce que *j'*, *t'*, *d'*, *qu'*, etc., existent nous écrivons *j'suis*, *t'arrives* ce qui entraîne une description morphologique plus proche de la réalité puisque l'opposition des clitiques sujet et objet de 2e personne est très affaiblie (on dit *t'arrives* comme *il t'a dit* et non *tu arrives*). Nous avons aussi noté deux degrés de « pauses », (au sens d'une perception subjective de pause, que celle-ci corresponde à un silence ou à un contour intonatif descendant).

11 Le corpus est donc exploitable pour des spécialistes de phonétique et de prosodie.

12 http://trans.sourceforge.net/en/presentation.php

(Barras *et al*. 1998) ; on peut réutiliser ces données dans le logiciel Transcriber pour rétablir l'alignement « fichier de transcription et fichier audio ». Les fichiers de transcription et les fichiers audio sont aussi directement accessibles en ligne de manière synchrone : la lecture de chaque alignement se faisant directement en ligne dans le navigateur de l'utilisateur, on utilise pour cela des ressources mises en œuvre par le CRDO[13]. Les fichiers de transcription ont été reformatés pour être disponibles en ligne dans un concordancier : les requêtes peuvent être faites sur les données brutes (le texte des transcriptions) ou sur une version étiquetée (*via* Cordial) des transcriptions. L'ensemble des fichiers de transcription est enfin disponible au téléchargement dans un format compatible avec les logiciels Lexico3[14] et Le Trameur[15] : il est ainsi possible de réaliser des opérations textométriques sur ces données.

Les ressources du corpus sont associées à des métadonnées qui permettent de définir un ensemble de descripteurs ainsi que les valeurs utilisées pour décrire et catégoriser le contenu et les objets des données visées. En règle générale, les métadonnées sont utilisées pour des opérations de recherche dans les données (sujet, auteur, date, mots clé, etc.). Par la suite, on pourra concevoir des opérations complexes permettant de croiser des requêtes et d'interroger simultanément les métadonnées et les contenus : on pourrait par exemple envisager de récupérer les ressources du corpus correspondant à une interrogation du type : « Quelles sont les fréquences d'utilisation du futur simple et du futur périphrastique chez les personnes interrogées de moins de 30 ans ? ». Concrètement, les métadonnées permettraient d'identifier les entretiens concernés par le critère d'âge, les identifiants des ressources résultantes devant permettre de calculer les fréquences visées (à condition qu'une annotation morphosyntaxique soit disponible sur les données du corpus).

Nous avons choisi de nous inscrire dans la démarche mise en œuvre par le CRDO pour mettre au point ces méta-

13 http://crdo.risc.cnrs.fr/exist/crdo/

14 http://www.tal.univ-paris3.fr/lexico/

15 http://www.tal.univ-paris3.fr/trameur/

descriptions des ressources du corpus. Les métadonnées construites sont au format OLAC[16] / Dublin Core[17]. La démarche suivie pour construire ces métadonnées a été la suivante : (1) rédaction « manuelle » d'une fiche descriptive de chaque interview (disponible en ligne au format PDF) ; (2) création d'un fichier de métadonnées pour chaque entretien à partir des informations disponibles dans les fiches initiales. Ces métadonnées ont été construites avec le logiciel makeMetadata[18]. Chaque entretien du corpus est associé *in fine* à une métadescription au sein d'un document XML, ces métadonnées étant codées avec des étiquettes Dublin-Core suivant les spécifications préconisées par OLAC. L'ensemble des fichiers de métadonnées constitue ainsi un catalogue de descriptions de l'ensemble des ressources du corpus, que l'on peut interroger pour mettre au jour des parties du corpus ayant telle propriété. On peut visualiser les métadescriptions de chaque entretien (en sélectionnant la propriété idoine dans le catalogue des métadonnées). Il est possible également de mettre au jour les métadescriptions contenant des mots-clés et accéder aux ressources associées (utilisation des métadonnées sélectives pour retourner au corpus).

3.2 L'accessibilité des données

Les ressources du corpus sont directement accessibles en ligne. Ce choix reflète notre conviction du caractère crucial que revêtent la constitution et la diffusion de « grands » corpus pour le français. Il est clair que des équipes universitaires repliées sur elles-mêmes ne parviendront pas à réaliser cet objectif et qu'il faut soutenir tout ce qui pourra à terme faciliter la circulation des ressources, les rendre utiles au maximum de chercheurs y compris lorsqu'ils travaillent sur des problématiques éloignées de celles qui sont à l'origine de CFPP2000. Une fois les enregistrements anonymés et transcrits, ils sont donc diffusés en ligne sans restriction[19]. Plus précisément, le corpus CFPP2000

16 Open Language Archives Community (http://www.language-archives.org).

17 Site web du DMCI : http://dublincore.org/

18 http://pi-ed268.univ-paris3.fr/MKM-doc/mkMETADATA.pdf

19 Les chercheurs à l'origine du projet demandent, en échange du travail de recueil, transcription du corpus et maintenance du site, qu'un article de présentation soit cité.

(fichiers audio et transcriptions) est disponible sous contrat *Creative Commons*[20].

4. Utilisations actuelles et développements futurs

Avec ses limites, le corpus a déjà donné lieu à des utilisations variées tant pour la langue que pour le discours.

Le français recueilli jusqu'ici est un français *commun* qui s'oppose, pour certains locuteurs au moins, aux vernaculaires de leurs *communautés* d'appartenance. Pour autant, le parisien urbain *commun* n'est pas le français *standardisé* décrit dans les grammaires, ce que montre la lecture des premières transcriptions. De plus, les discours recueillis ne sont pas homogènes. Ils varient en fonction des caractères sociologiques des enquêtés, (mais tout autant en fonction de leurs trajectoires personnelles complexes). Ils varient aussi au cours d'un même entretien en fonction du lien établi avec l'enquêteur et des thèmes abordés.

4.1 Des spécialistes de l'oral, prosodistes et syntacticiens

Le découpage en unités syntaxiques a retenu l'attention des syntacticiens qui se sont appuyés sur CFPP2000 pour leurs recherches. Des structures particulières ont commencé à être abordées, comme celle des constructions binaires à prédicat focalisé (Tanguy, à paraître). F. Lefeuvre (2010, à paraître) examine les marqueurs *bon* et *quoi* dans leur délimitation de segments verbaux et averbaux et C. Détrie (2010) envisage les particules « tu vois, vous voyez ». Une confrontation avec d'autres corpus de données orales a pu aussi être établie : ainsi F. Lefeuvre, M.-A. Morel et S. Teston-Bonnard (2010) étudient *quoi* dans trois types différents de corpus oraux : CFPP2000, Clapi, et les corpus de l'EA 1483 (Paris 3). Un numéro de *Langue française* coordonné par F. Lefeuvre et E. Moline (« Unités syntaxiques et unités prosodiques ») rassemble des articles dans lesquels les auteurs s'interrogent sur la congruence entre unités syntaxiques et unités prosodiques. Dans ce volume, ont été ainsi réunis quelques-uns des spécialistes du domaine, à qui il a été demandé de présenter leur méthode d'analyse, puis

20 http://creativecommons.org/licenses/by-sa/2.0/fr/deed.fr

de l'appliquer à un même extrait oral, [11-01], tiré de CFPP2000. Quatre types de réponse ont été apportés : certains auteurs privilégient l'apport de la syntaxe, (i) soit en ayant exclusivement recours à des unités de type syntaxiques (P. Le Goffic), soit en postulant que les constituants prosodiques sont fondés sur des bases syntaxiques (Delais-Roussarie, Post et Yoo). (ii) Ou bien, la congruence entre les unités syntaxiques et les unités prosodiques n'est pas considérée comme aussi fréquente (cf. Simon & Degand ; Lacheret *et al*.), ce qui peut conduire à la recherche de nouvelles unités de segmentation, comme les unités de base de discours chez Simon & Degand. Quoi qu'il en soit, la rection constitue pour tous ces auteurs le point d'ancrage de l'identification des unités. (iii) Ou bien encore la délimitation des unités s'effectue par l'association de marques prosodiques et de marques textuelles ou informationnelles (Berrendonner, Cresti *et al*). (iv) Enfin, M.-A. Morel d'une part, C. Blanche-Benveniste et Ph. Martin d'autre part, accordent une part prépondérante à la prosodie, qu'il s'agisse de l'identification des unités (Morel) ou de l'interprétation du texte (Blanche-Benveniste & Martin).

4.2 En sociolinguistique

Des sociolinguistes et des analystes du discours ont commencé à s'intéresser aux corrélations entre formes linguistiques et représentations sociales.

Les travaux menés jusqu'ici débouchent d'abord sur une compréhension accrue de la place qu'occupe l'individu dans l'ensemble, par définition hétérogène, que constitue une variété sociale. S. Fleury et S. Branca-Rosoff (2010) ont abordé ainsi l'alternance futur simple / futur périphrastique dont la répartition obéit à des contraintes linguistiques (telle que l'affinité du futur simple et de la négation) plutôt qu'à des contraintes sociolinguistiques qui indiqueraient la prochaine disparition du futur simple ; ils remarquent aussi l'incidence très forte des variations individuelles sur les fréquences d'usage des formes, en particulier quand il s'agit de routines communicatives à peu près inconscientes (*je vais dire X… je dirai X*).

Le corpus peut également être exploré de façon qualitative. Les analyses remettent ainsi en cause, pour certains

locuteurs, les renseignements fournis dans les métadonnées et conduisent à récuser les appartenances calculées uniquement à partir de données préalables (niveau d'étude, ou catégorie socioprofessionnelle) et à s'intéresser à la pertinence d'appartenances revendiquées. Selon une première perspective, le chercheur part du texte. Les interviews sont concaténées en un seul document d'où sont extraits des énoncés, afin d'étudier les propriétés syntaxiques et énonciatives d'une variable. Selon une seconde perspective, complémentaire, le chercheur revient à des entretiens individualisés considérés comme des totalités spécifiques et il envisage les variables comme autant de points de condensation exprimant des identités discursives. Le fait que les entretiens fassent une place importante aux discours des locuteurs sur leurs appartenances identitaires permet de montrer l'articulation des comportements linguistiques et des positionnements représentés. Le travail de J. M. Barbéris (2010) sur l'emploi générique de la 2e personne incarne exemplairement cette double direction de travail[21].

Il est possible aussi d'aborder la question en partant des représentations des locuteurs. Ce chantier a été récemment ouvert par M. Pires et S. Branca qui vont exploiter les réponses aux questions sur l'identité et sur « le mauvais français » pour tracer des « portraits sociologiques »[22] des locuteurs. Ces travaux sur les normes et sur les identités revendiquées (par opposition aux normativités constatées dans l'interaction) recoupent l'analyse menée par J.-M. Barbéris et devraient mieux faire apparaître la complexité des positionnements sociaux presque toujours non homogènes. Par exemple, l'identification possible à une profession, l'appartenance à une certaine partie de la ville, ou à des micro-communautés fondées sur des loisirs partagés, aboutissent à des goûts et des dégoûts qui ne découlent pas automatiquement des catégories socio-professionnelles. Cela ne veut pas dire que toute généralisation est impossible, mais qu'il

21 Nous avons aussi essayé de montrer que dans ce corpus la signification du mot *village* est inséparable de l'usage social de ce mot et qu'en comprendre la logique demande de prendre en compte le contexte urbain et la période « post-moderne ». (Branca 2010 ; voir aussi Barbéris 2010).

22 Voir Lahire 2002.

faudra sans doute travailler davantage avec les notions de réseaux sociaux.

4.3 Les perspectives. L'extension du corpus

Le projet a été initié grâce à une aide de 10 000 euros obtenue dans le cadre du projet de la ville de Paris « Dynamiques de l'agglomération parisienne » ; il se poursuit grâce au soutien de la DGLFLF. Cependant, en 2011, CFPP2000 reste pour l'essentiel une réalisation fondée sur le bénévolat et sur des heures de vacations pour des transcripteurs en situation précaire. Dans l'état actuel, l'échantillonnage n'est pas suffisamment équilibré et la deuxième étape devra améliorer ce point. Convaincre les locuteurs de diffuser en ligne leur enregistrement s'est révélé souvent difficile. Les enquêtés ont été approchés grâce à des « réseaux sociaux ». Nous avons eu souvent recours à « l'ami d'un ami » (Boissevain 1974). La confiance ainsi obtenue a permis de lever les difficultés, mais a ralenti le travail de collecte dans les milieux éloignés du microcosme universitaire. Les données sont d'ailleurs en nombre trop limité pour aborder de manière détaillée l'ensemble des paramètres commandant la distribution morphosyntaxique (exemple du futur et des affinités entre lexique verbal et choix des temps). Cependant, le corpus n'est pas clos : il a vocation à être complété dans les années qui viennent de façon à mieux représenter la diversité de la population ; en particulier, il fera de la place aux habitants des quartiers de l'ouest et aux habitants des banlieues et il intégrera quelques locuteurs plurilingues : locuteurs parlant une langue romane (espagnol), slave (serbe) ou appartenant à des familles non-indoeuropéennes (chinois et arabe).

Nous travaillons avec des chercheurs de Valibel et de l'ATILF et de l'ANR Rhapsodie à un projet de description plus fine des métadonnées nécessaires pour les corpus oraux et sur la possibilité de faire des requêtes permettant de réaliser des extractions en fonction de paramètres articulant sociologie et formes linguistiques. L'amélioration de la description des ressources du corpus permettra d'échanger, de partager ces ressources avec d'autres centres et de les archiver dans de grandes banques de données de français parlé.

Enfin, les interviews sont des images sonores du Grand Paris qui s'adressent non seulement aux chercheurs, mais aussi aux curieux, aux flâneurs et à tous ceux, Parisiens ou visiteurs, qui s'intéressent au patchwork des lieux qu'ils traversent ou habitent, et aux réseaux sociaux dans lesquels des milliers de citadins interagissent, construisant ainsi jour à jour le sens de la ville. A côté de la ville monumentale, l'auditeur découvre les espaces partagés du quotidien, rues commerçantes, jardins, cafés, marchés, terrains de foot et d'autres lieux plus secrets tel ce ponton à la limite du Port de l'Arsenal où l'on bronze au soleil en se rêvant très loin, ou ce restaurant d'habitués où le patron vous garde votre serviette. Dans ce Grand Paris mosaïque, la plupart des migrants refusent de se laisser enfermer dans une identité, et revendiquent des appartenances multiples. Même les plus nostalgiques, qui de la petite Italie banlieusarde des années 60, qui des communautés algériennes où tout le monde se connaissait, sont en même temps ancrés dans un présent où les sociabilités passent par les réseaux sociaux et la pratique des sports collectifs. On perçoit aussi la variabilité des oppositions qui paraissent les plus stables : dans les banlieues populaires de Saint-Ouen ou de Rosny, les jeunes gens se construisent contre la capitale bourgeoise qui les repousse, mais ils racontent avec plaisir leurs incursions dans le Paris monumental. Quant aux habitants du centre-ville et des banlieues bourgeoises de l'Ouest, la plupart s'aventurent peu dans les quartiers populaires de la périphérie, mais même alors, ils ne sont pas totalement indifférents au foisonnement et à la vitalité d'une ville multiple.

Références bibliographiques

Barbéris J.-M. (2010). « "Quand t'es super bobo"… La deuxième personne générique dans le français parisien des jeunes », cmlf/2010258.

Blanche-Benveniste C. & Jeanjean C. (1987). *Le français parlé. Transcription et édition*. Paris : Didier Érudition.

Blanche-Benveniste C. (2000). « Convergences de matériel grammatical permettant d'établir des typologies textuelles », *in* M. Bilger (coord.) *Linguistique sur corpus.*

Études et réflexions 31. Perpignan : Presses universitaires de Perpignan, 103-116.

Boissevain J. (1974). *Friends of friends : Networks, manipulators and coalitions*. Oxford : Basil Blackwell.

Branca-Rosoff S. (1999), « Types, modes et genres entre langue et discours », *Langage et Société* 87 : 5-24.

Branca-Rosoff S. (2010). « La nomination des lieux et des habitants de la ville et la référence à un univers de discours 'autre' dans un corpus d'interviews non directives » *Colloque international Dialogisme : langue, discours*, Praxiling, 8-10 septembre 2010.

Branca-Rosoff S., Fleury S., Lefeuvre F. & Pires M. (2008). *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)* (http://cfpp2000.univ-paris3.fr/).

Bulot T. (dir.) (2004). *Les parlers jeunes* (Pratiques urbaines et sociales). *Cahiers de Sociolinguistique* 9. Rennes : Presses Universitaires de Rennes, 176 pages.

Détrie C. (2010). « De *voir* à *tu vois* / *vous voyez* : fonction sémantico-énonciative et postures énonciatives construites par ces particules interpersonnelles », DOI : 10.1051/cmlf/.

Dister A., Francard M., Geron G., Giroul V., Hambye P., Simon A.-C. & Wilmet R. (2006). *Conventions de transcription régissant les corpus de la banque de données* VALIBEL http://valibel.fltr.ucl.ac.be, corpus.

Fleury S. & Branca-Rosoff S. (2010). « Une expérience de collaboration entre linguiste et spécialiste de TAL : L'exploitation du corpus CFPP 2000 en vue d'un travail sur l'alternance Futur simple / Futur périphrastique », *CAFLS* 16(1) : 63-98.
http://cfpp2000.univ-paris3.fr/CFPP2000.pdf

Kerswill P. & Cheshire J. (s. d.). *Linguistic innovators : The English of Adolescents in London.*
www.lancs.ac.uk/sss/project/linguistics/innovators/index.htm

Lahire B. (1998). *L'homme pluriel. Les ressorts de l'action.* Paris : Nathan.

Lefeuvre F. (à paraître). « *Bon* et *quoi* à l'oral : des marqueurs d'unités syntaxiques averbales autonomes », *Linx* (Krazem éd.).

Lefeuvre F. (2010). « *Bon* à l'oral : une unité syntaxique averbale et autonome ? », *in* I. Behr et F. Lefeuvre (éd). Paris : Ophrys, 124-136.

Lefeuvre F., Morel M.-A. & Teston-Bonnard S. (à paraître). « Valeur prototypique de *quoi* à travers ses usages en français oral et contemporain », *Neuphilologische Mitteilungen* (Bulletin de la Société Néophilologique).

Lodge A. (2004). *A Sociolinguistic History of Parisian French*. Cambridge : Cambridge University Press.

Mellet S. (2002). « Corpus et recherches linguistiques », *Corpus* [En ligne], n°1 | novembre 2002, mis en ligne le 15 décembre 2003, Consulté le 27 octobre 2010. URL : http://corpus.revues.org/index7.html

Nicolaï R. (2000). *La traversée de l'empirique*. Paris : Ophrys (« Bibliothèque de Faits de Langues »).

Sankoff D. & Laberge D. (1978). « The Linguistic Market and the Statistical Explanation of Variability », *in* D. Sankoff (ed.), *Linguistic Variation : Models and Methods*. New-York : Academic Press, 239-250.

Tanguy N. (2010). « Étude des compléments différés à l'oral à l'interface syntaxe – prosodie ».
CMLF : http://dx.doi.org/10.1051/cmlf/2010190.

Tanguy N. (à paraître). « Complémentations en direct. Le fonctionnement des compléments différés à l'oral », *in* Actes du Colloque « Complémentations », 20-23 octobre 2010, Universidade Santiago De Compostela, Espagne.