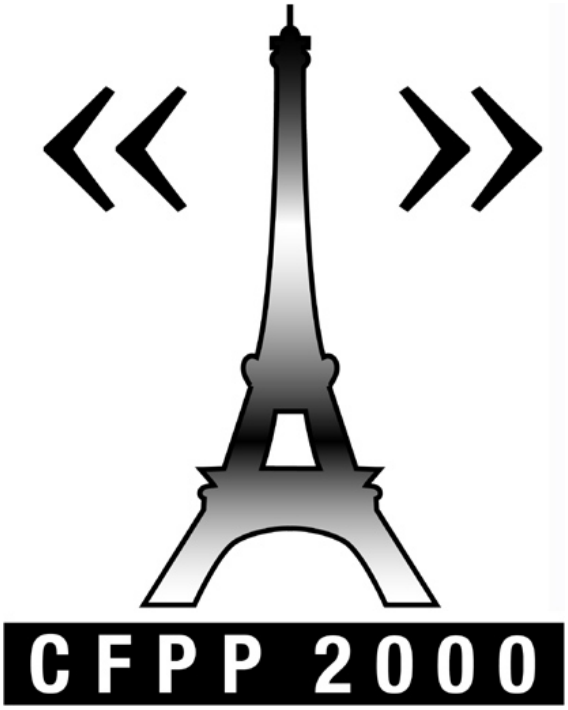
Université Sorbonne Nouvelle Paris 3

CFPP2000



Discours sur la ville

Corpus de Français Parlé Parisien des années 2000

Branca-Rosoff S., Fleury S., Lefeuvre F., Pires M.

2012

Discours sur la ville

Corpus du Français Parlé Parisien des années 2000

CFPP2000

Convaincue que le développement des technologies de la parole est un enjeu majeur pour les sciences humaines, une petite équipe de l'université de Paris 3-Sorbonne-Nouvelle (Clesthia-Syled, EA 2290), constituée à l'initiative de Sonia Branca-Rosoff, a entrepris de rassembler des interviews longues, réalisées auprès d'habitants de Paris et de sa périphérie, sur la vie des habitants et sur les rapports qu'ils entretiennent à leurs quartiers, afin de développer un corpus destiné à des études sur les pratiques linguistiques et les représentations discursives. L'époque actuelle est propice à de telles entreprises car le développement d'Internet et l'existence de logiciels libres facilitent considérablement la recherche de solutions de transcription, de stockage et de partage des données. Il est aussi plus facile d'outiller les corpus et d'en permettre une première exploration en ligne, grâce notamment à des concordanciers et aux progrès des programmes de lexicométrie. L'équipe pouvait donc contribuer aux efforts entrepris à partir des années 2000 pour rapprocher la France des critères internationaux en matière de grands corpus linguistiques. Outre S. Branca-Rosoff, elle comprend Florence Lefeuvre et Serge Fleury, ce dernier assurant l'organisation du site et le développement d'outils informatiques, ainsi que Mat Pires, de l'université de Franche-Comté (équipe Elliadd EA 4661).

Ce Corpus de Français Parlé Parisien (désormais noté CFPP2000) a commencé à être recueilli en 2005-2006 grâce à l'appel d'offres de la Mairie de Paris, "Dynamiques de l'agglomération parisienne", qui a fourni les ressources indispensables à l'achat de matériel et au financement des premières transcriptions. Le projet s'est poursuivi avec quelques modifications grâce à un soutien de la DGLFLF, modeste mais précieux, jusqu'à la réalisation d'un site, alignant données sonores et transcriptions : http://cfpp2000.univ-paris3.fr. En mars 2012, le corpus comportait 535 000 mots environ, soit 37h 75. pour 29 entretiens (dialogues ou multilogues). Les transcriptions ont été mises en ligne avec une première phase d'outillage des données disponibles, notamment des concordanciers[1]. Les métadonnées qui accompagnent le corpus permettent de sélectionner des entretiens correspondant aux critères visés (par exemple, sélectionner les entretiens sur des critères d'âge). La collecte des données doit se prolonger dans les années à venir jusqu'à atteindre un million de mots. Une seconde étape d'outillage vise à étendre les fonctionnalités déjà disponibles en améliorant métadonnées et annotations.

CFPP2000 est caractérisé par ses objectifs : permettre une approche du français de Paris et de sa banlieue limitrophe, ou de façon plus précise permettre – dans son unité comme dans sa variation – l'étude du français « de communication » que les participants adoptent lors d'interviews conversationnelles[2]. L'accommodement réalisé entre partenaires aboutit à une variété que l'on peut opposer aux vernaculaires dont certains participants usent dans leurs communautés d'appartenance. Cependant, ce français « commun » ne se confond pas avec le standard parce qu'il relève d'une oralité ordinaire encore mal prise en compte par les grammaires et qu'il contient des façons de dire que grammaires et dictionnaires considèrent comme familières. Ce n'est pas non plus un bloc homogène ; il inclut des variations, en fonction des appartenances sociales des habitants interrogés, des activités discursives des interlocuteurs et des moments de la conversation.

Les métadonnées associées au corpus donnent la possibilité d'opérer des tris qui tiennent compte du lieu d'habitation, de l'âge, du niveau d'étude, des catégories socioprofessionnelles des interviewés ; les activités linguistiques et les thématiques comparables donnent du sens à des comparaisons. Enfin la durée moyenne des interviews qui avoisine une heure rend possibles des va-et-vient entre des approches quantitatives et des retours à une analyse qualitative tenant compte des particularités des locuteurs et de leurs positionnements au cours de l'entrevue.

Le système de transcription adopté se prête à des études sur le discours, sur la variation sociolinguistique, sur la syntaxe du français parlé. Pour des recherches portant sur la phonétique, sur l'intonation ou sur les interactions, il est toujours possible de revenir au son, puisque les transcriptions

[1] Les concordanciers sont des programmes qui permettent de repérer toutes les occurrences d'une forme avec ses contextes (CFPP2000 comporte des outils de recherche correspondant à des séquences textuelles, à des catégories étiquetées morphologiquement, etc.)

[2] Interviews dont les participants savent qu'elles seront mises en ligne et rendues ainsi accessibles dans l'espace public.

ont été alignées sur le son au tour de parole. Dans les années qui viennent, nous ajouterons un alignement au phonème (*EasyAlign* puis correction manuelle).

Une fois la bande-son anonymisée, les données sont mises progressivement en ligne et rendues accessibles sans restriction à toute personne intéressée, chercheur en sciences humaines, ou tout simplement curieux amoureux de Paris, désireux d'entendre des habitants parler de leur vie dans la capitale.

1. Le choix des données : le genre de l'entretien, Paris et les communes limitrophes comme terrain, les locuteurs et la question de l'échantillonnage

Il ne peut exister d'échantillon représentatif de « tous les discours en circulation ». Un corpus, s'il est exploitable, est toujours de façon plus modeste le résultat d'un choix, qui suppose toujours une certaine idéalisation des locuteurs regroupés en catégories et de leurs activités linguistiques décrites en termes de genres discursifs. Il est donc important d'expliciter nos options, même si nous avons cherché à constituer un corpus qui puisse se prêter à différentes sortes d'utilisation.

1.1. Des entretiens longs, semi-directifs, sur le rapport des Parisiens à leur quartier

(i) Le questionnaire sous-jacent aux entretiens est déterminant si l'on veut pouvoir rapprocher des données différentes et comparables. Nous reconnaissons, comme C. Blanche Benveniste (1997), que les phénomènes de variation liés aux usages différents de la langue sont plus contrastés que les phénomènes relevant des différences sociales et qu'on ne saurait comparer sans absurdité des débats menés par des hommes politiques à la radio avec le récit de l'arrivée en France d'un immigré. Pour autant la question de l'impact des variations diastratiques n'est pas tranchée puisqu'on ne dispose pas d'un corpus important permettant d'observer dans des activités du même type les comportements des locuteurs ayant, par exemple, suivi une scolarité longue par rapport aux locuteurs ayant interrompu leurs études. C'est d'abord cette question complexe que l'entretien semi-directif permet de poser sur des bases empiriques. Il faut cependant noter que ce genre est toujours un compromis entre des exigences contradictoires : assurer la permanence des thématiques pour pouvoir ensuite traiter les productions des locuteurs comme plus ou moins équivalentes ; inversement, laisser de l'initiative aux interviewés pour que leurs conceptions puissent émerger, et en même temps rester au plus près du ton d'une conversation naturelle en évitant les longs silences ou les relances caractéristiques des entretiens cliniques (« mm », répétition des derniers mots, etc.).

Sans pouvoir supprimer les contradictions entre ces buts divergents, on peut essayer de les atténuer : d'abord et avant tout en s'intéressant aux opinions des interviewés, en leur manifestant l'importance qu'a leur témoignage, en nouant une relation de confiance avec eux ; ensuite, en leur laissant le temps de construire leurs réponses, de donner leurs raisons de penser comme ils le font. Enfin, en n'hésitant pas à confronter les enquêtés à des points de vue différents du leur comme cela se produit au cours de conversations banales. La fameuse exigence de neutralité que rappelle chaque manuel du parfait enquêteur a paru parfois moins utile que l'échange naturel des opinions, même si l'enquêteur a eu soin de ne pas heurter trop ses partenaires[3].

[3] Cette liberté par rapport au questionnaire a pour conséquence que l'enquêteur pose parfois des questions bizarres, ou dérangeantes. Il est difficile de contrôler ses prises de paroles en essayant de mener une conversation « naturelle », tout en cherchant parfois à déclencher des activités linguistiques précises. Ces ruptures n'ont pas paru trop graves.

(ii) L'entretien favorise de longues prises de parole.

La durée de l'échange est apparue comme un élément important du projet : la relation d'enquête pousse les enquêtés à se rapprocher par un effet mimétique automatique, bien connu des sociolinguistes, de la variété employée par l'enquêteur, les deux parties pratiquant plus ou moins consciemment un français qui tend vers une langue *commune*. Toutefois, et pour peu que l'enquêteur adopte un style détendu, le déroulement de l'entretien favorise l'évolution des échanges d'un registre relativement formel vers un registre plus « vernaculaire ».

De longs corpus augmentent aussi les chances de rencontrer plusieurs attestations d'une même tournure, ce qui caractérise certains des idiolectes (ainsi les 9 occurrences de *en revanche* pour la première tranche d'entretiens comportant 300 000 mots environ se rencontrent seulement chez Yvette Audin et une des enquêtrices (*cf.* Branca-Rosoff *et al.* 2009 en ligne).

Enfin de longs corpus permettent d'opérer un va-et-vient de l'identification minimale fournie par les paramètres sociologiques, âge, niveau d'étude, profession, etc., à l'auto-description complexe que les habitants font d'eux-mêmes et des façons dont ils s'inscrivent dans leurs lieux d'habitation.

(iii) l'entretien est un discours *sur* l'action et pas un discours *dans* l'action.[4]

L'entretien incite à un retour réflexif sur ses pratiques, propice au développement d'une syntaxe complexe. Il fournit ainsi des données complémentaires aux données que l'on recueille lorsqu'on collecte du langage dans l'action.

Dans le sillage de l'analyse conversationnelle d'inspiration ethnométhodologique, qui donne le primat au lien entre parole et activités (conversationnelles ou non), certains chercheurs accordent une place seconde à l'inventaire des formes du langage préférant se centrer sur les procédures par lesquelles les interlocuteurs mobilisent telle ou telle de ces formes au cours de leurs activités langagières. Le corpus CFPP2000 traite uniquement des échanges en situation d'entretien : par exemple, il ne comporte pas d'enregistrements sur les lieux de travail, dans les commerces ou dans la famille. Plus généralement, les façons de faire et de dire des locuteurs ne sont pas *observées* par le chercheur. Elles sont *racontées* par les locuteurs qui en proposent une *représentation.*

Toutefois, l'entretien conversationnel constitue lui-même un des usages de la langue bien repérable dans nos sociétés. Il peut être étudié comme tel, en tant qu'activité langagière, et ce type d'interaction est particulièrement pertinent quand il s'agit de travailler sur le rapport aux normes sociales puisque les enquêteurs universitaires qui interagissent avec les enquêtés, peuvent (selon le rôle qu'ils s'attribuent et qu'on leur attribue au cours de l'interaction) représenter une figure normative ou non.

Par ailleurs, l'entretien présente des avantages lorsqu'on s'intéresse aux formes de la langue : outre l'homogénéité relative du matériau collecté, cette situation permet à l'enquêteur de provoquer l'emploi d'une certaine variété de formes selon qu'il sollicite des récits, des descriptions ou des justifications (voir 1.2). Si l'on pense que ces ressources linguistiques sont relativement stables[5], c'est-à-dire qu'elles préexistent à leur usage, que les locuteurs ne les « réinventent » pas constamment dans l'interaction, l'entretien présente l'intérêt de susciter l'apparition de formes variées et de ne pas se limiter aux routines suffisantes pour l'action quotidienne.

(iv) L'entretien permet d'introduire dans la conversation un ensemble de thématiques urbaines.

- L'intérêt de ce questionnement sur leur pratique de la ville est apparu immédiatement aux informateurs qui ressentent tous qu'ils ont une expérience à transmettre.

- D'autre part, le thème de la ville facilite le positionnement des interlocuteurs pendant l'interview. Il est en effet plus facile d'accepter de répondre à des questions sur son cadre de vie que d'accepter un entretien visant à observer la complexité de ses pratiques langagières ou à situer ses productions langagières selon une stratification hiérarchique allant du plus familier (ou du plus populaire), au plus surveillé (ou au plus bourgeois). Il est encore plus difficile, même si l'on accepte cette situation inconfortable, de s'exprimer de façon libre et spontanée (voir les réflexions de Labov 1976 qui a insisté

[4] Voir Renaud 2004.

[5] Pour une opinion contraire, Mondada 1999.

sur la contradiction entre la situation artificielle créée par la présence d'un observateur et le souci de recueillir le parler vernaculaire des locuteurs).

- Des motifs plus centraux ont également joué dans le choix de la thématique de la ville : la dimension urbaine est une dimension centrale de l'expérience sociale en ce début du XXI^{e} siècle où plus de 80% de la population française vit en ville. Un des buts des entretiens collectés est d'interroger les habitants sur leurs représentations de Paris (espace englobant et unifiant ou espace constitué d'un conglomérat de territoires fragmentés, voire antagonistes) et sur la façon dont ils pensent leurs rapports aux différents groupes qui peuplent la ville (voisins, condisciples, collègues) (voir notamment Bulot 2003, 2004, 2005, Mondada 2000). On peut faire l'hypothèse que ces représentations sont en rapport avec la différenciation linguistique. L'influence des territoires de la ville peut passer par l'identification à un territoire nettement délimité, que l'on ait affaire à des lieux d'intense brassage de populations, comme c'est le cas pour les communes de Saint-Ouen ou de Pantin, ou au contraire à des quartiers relativement homogènes comme l'a été longtemps le 7^{e} arrondissement de Paris. Des variétés urbaines peuvent aussi émerger par le biais des réseaux de socialisation (Milroy & Milroy 1992), paroisses, lycées, groupes de jeunes, habitués de certains cafés, etc. Les phénomènes remarquables peuvent concerner de façon étroite l'adoption de telle ou telle variante, mais aussi l'usage de registres spécifiques plus ou moins consciemment adoptés en fonction de la position du quartier dans l'espace urbain. Une habitante du XI^{e} développe ainsi le thème du rôle structurant de l'opposition est/ouest à Paris. L'opposition qu'elle établit entre les bobos et les bourgeois renvoie selon elle à des thèmes de conversation, mais aussi à un « style » relâché opposé au style « tendu » des habitants de l'ouest de Paris. Ses représentations sont inséparablement des représentations d'un territoire, du groupe qui l'habite et de comportements qui structurent des processus identitaires différenciateurs.

(v) Les thématiques abordées

Au-delà d'un matériau suffisamment homogène et qui convienne aux linguistes, l'entretien thématique a permis de récolter un matériau précieux pour les historiens, les sociologues, les analystes du discours qui pourront utiliser ces données dans la perspective de leurs disciplines respectives.

Il comporte des questions sur les réactions des habitants face aux changements intervenus dans la composition de la population, dans l'urbanisme, dans les commerces, etc. Il interroge les habitants des quartiers sur les effets de « la mondialisation » au plan des mélanges de population, des modes culinaires ou vestimentaires. Il leur demande d'évoquer la façon dont ils vivent des situations de « mixité » ou de « ségrégation ». Accessoirement, le thème de la ville a permis d'intégrer des questions sur les pratiques plurilingues et sur les traces qu'elles laissent sur le français, qu'il s'agisse des mots provenant de populations historiquement installées depuis longtemps comme les Manouches de Montreuil ou encore des argots professionnels comme l'argot des chiffonniers de Saint-Ouen ou qu'il s'agisse de la présence récente du multilinguisme européen, africain, asiatique. Le questionnaire permet aux citadins de faire le point sur leurs sentiments par rapport à toutes ces langues : essaient-ils de les maintenir et de les transmettre si leurs parents parlent une autre langue que le français, de se les approprier s'il s'agit de leurs voisins ? Des questions sur les parlers des jeunes les amènent à commenter leur stigmatisation éventuelle et à s'exprimer sur le degré de prestige qu'ils attribuent au standard.

Ce questionnaire porte encore sur la relation au « temps social » ; le développement de nouvelles formes de fêtes est souvent évoqué dans cette rubrique, fêtes de voisins (d'immeubles), de quartier, nuit blanche, fête de la musique.

L'espace urbain fait également l'objet de questions portant d'une part sur les clivages traditionnels, l'opposition rive gauche, rive droite ; Paris est/ouest, Paris centre/banlieue ; d'autre part, sur les problèmes de déplacements en ville, ce qui conduit les enquêtés à se remémorer l'espace où ils vivent, ainsi que leurs déplacements et à en proposer une expression verbale.

Enfin, l'action politique des municipalités (et de l'État) est présente dans beaucoup d'entretiens, souvent introduite par des questions sur la migration, sur l'école, sur la grande pauvreté ou sur la façon de s'informer sur le quartier.

1.2. Des entretiens qui présentent deux types de contrastes

Si le genre entretien est caractérisé par les rôles différents des partenaires (l'enquêteur restant celui qui pose les questions) et par le travail de retour sur soi demandé à l'enquêté, il peut s'accompagner de variations à la fois sur le plan interne et sur le plan du dispositif d'enquête.

(i) Des types de discours

À la variation des registres déjà évoquée en 1.1 (ii), s'ajoute la variation induite par des questions destinées à susciter diverses activités linguistiques. Les locuteurs sont invités à *décrire* leurs trajets dans la ville, puis à *raconter* des anecdotes frappantes liées au quartier, à donner leur opinion *argumentée* sur des questions plus ou moins polémiques, à *s'exprimer de façon métalinguistique* sur les langues, particulièrement sur le français.

Voici à titre d'exemple un fragment narratif :

> 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 + +
> [CFPP2000 12-03, Valentine Testanier, femme, 60 ans, concierge]

Voici une brève description d'itinéraire du domicile au travail :

> alors oui je sors de chez moi ben je prends en fait là le Mail c'qu'on appelle le Mail du Centre Ville euh + après on arrive à la Place euh de l'Eglise + (oui) et Place Carnot où y a le marché et après ben je remonte euh + + où y a la piscine et j'arrive au Mail ben au Mail Jean-Pierre Timbault où c'est (mm) l'adresse de la CAF quoi
> [CFPP2000 RO-01, Isabelle Legrand, femme, 32ans, employée de la Caisse d'Allocations Familiales]

(ii) des dialogues et des polylogues

Les entretiens CFPP2000 partagent le même genre mais contrastent en ce qui concerne le nombre de participants. Certains sont des "polylogues" : la présence de deux personnes, ou plus, face à l'interviewer atténue l'impact d'un observateur extérieur et contribue à détendre l'atmosphère ; elle favorise la discussion, facilite l'expression des désaccords entre les participants ou l'élaboration commune d'opinions. Certains sont des interviews en face à face qui suscitent des prises de parole plus longues de l'enquêté, l'enquêteur se bornant souvent à l'accompagner par des marques d'approbation (« mm », « oui », etc.), sans l'interrompre. Ces interviews offrent un matériau plus facile à exploiter aux spécialistes de prosodie puisqu'elles contiennent peu de chevauchements. On peut donc contraster les interviews selon le nombre de participants (dialogues *vs* polylogues).

1.3. Les zones d'enquête : Paris et la proche banlieue

L'île de France accueille une population de 11,7 millions d'habitants, 19% de la population française, métropolitaine[6]. Son importance symbolique est encore plus grande : des études portant sur d'autres pays ou périodes ont montré que les variétés langagières pratiquées dans les très grandes villes ont un prestige social qui en font des moteurs du changement linguistique (*cf*. pour la Grande-Bretagne, les travaux de Kerswill & Cheshire sd et Torgersen 2006). Et pour ce qui est de Paris, A. Lodge (2004) a établi que les pratiques langagières de la capitale ont exercé depuis l'origine une influence décisive sur l'ensemble de la France. Paris pèse également d'un poids très lourd sur la francophonie, bien davantage que Madrid pour l'espagnol, Lisbonne pour le portugais ou Londres pour l'anglais. Dans ces dernières villes, les anciennes métropoles comptent à présent moins que Mexico, São Paulo ou New-York, alors que Paris a une position dominante parce que c'est la plus grande des villes où le français est la langue de la quasi-totalité de la population. Aussi, il est intéressant de se demander si les innovations

[6] Source : Insee, janvier 2009.

linguistiques qui y voient le jour seront adoptées dans le reste de l'espace francophone. Paradoxalement, Paris jusqu'ici n'a pas fait l'objet d'une enquête sociolinguistique dont les données soient facilement accessibles : le Corpus de Référence du Français Parlé (CRFP) élaboré par l'équipe Delic, qui comporte 439 000 mots, a été recueilli dans 37 villes de France ; Paris et sa région n'interviennent que pour 20% du total. Le corpus ESLO[7], élaboré sur des bases sociologiques, concerne une ville moyenne Orléans. Le corpus PFC (Phonologie du français contemporain) rassemble des enregistrements collectés dans toute la francophonie.

En constituant un corpus d'un million de mots sur le parisien d'Île de France, nous rendons possibles les comparaisons entre le français parisien et le français d'autres pays francophones (corpus Sankoff-Cedergren, 1971, Thibault & Vincent 1984, Hull-Ottawa, corpus du français parlé québécois de G. Dostie) pour le Canada ; corpus Valibel pour la Belgique).

1.4. Les problèmes d'échantillonnage

Paris et sa couronne forment à bien des égards une seule agglomération, même si les boulevards périphériques constituent une frontière visible et symboliquement importante. Nous avons donc pensé nécessaire d'inclure quelques communes limitrophes dans l'échantillon retenu pour mieux représenter l'ensemble complexe des variétés du français de la capitale.

En 2005, nous étions partis de six zones d'enquête, approximativement homogènes, en tenant compte à la fois des découpages administratifs (les résultats de l'enquête INSEE sont fournis par arrondissement) du nombre d'habitants, du poids moyen des cadres supérieurs et des patrons (plus de 70% de cadres dans les V^{e}, VI^{e} et VII^{e} arrondissements, moins de 25% dans le XX^{e}) et du prix moyen au m^{2} des appartements vendus libres[8]. Nous pensions alors contraster fortement les zones d'enquête, tout en tenant compte des représentations traditionnelles sur l'existence d'un sociolecte bourgeois propre à Neuilly, d'un parler parisien populaire de Belleville, d'une communauté linguistique émergente influencée par l'émigration à Montreuil. Pour chaque zone d'enquête, nous avions prévu d'interviewer dix-huit locuteurs (9 hommes et 9 femmes) répartis en trois catégories bien contrastées selon l'âge, afin de pouvoir repérer des innovations ou au contraire des formes vieillissantes (les 16-25 ans ; les 45 à 50 ans, groupe constituant pour nous le groupe des adultes ; enfin, le groupe des 60 ans et plus correspondant aux personnes âgées). Les catégories sociologiques retenues avaient été également ramenées à trois (à partir d'une pondération tenant compte du travail, du diplôme, et pour les jeunes du travail des parents). Pour chaque zone, il était prévu de retenir deux catégories sociologiques en fonction de l'habitat (par exemple, à Neuilly des classes supérieures et moyennes, à Montreuil, des classes moyennes et populaires).

Nous avions sous-estimé deux facteurs : enquêtes et transcriptions, qui supposent d'y consacrer de longues heures, ont été réalisées pour l'essentiel par des enseignants chercheurs, dans une période universitaire peu propice à l'activité bénévole. Soumis à des contrôles de plus en plus tatillons, les chercheurs sont conduits à surinvestir les aspects les plus visibles de leur travail au détriment du travail concret. Personne dans la petite équipe de départ n'a pu consacrer des journées au quadrillage systématique du terrain. De plus, de nombreux locuteurs, d'abord intéressés par le projet, ont reculé lorsqu'ils ont appris que ce n'était pas seulement le contenu de leurs propos qui serait rendu public, mais leur discours tel quel avec ses répétitions, ses lapsus, ses énoncés rompus, ses « bon » et ses « ben » phatiques. Ils ont surtout été intimidés par l'idée qu'on pouvait reconnaître leur voix, leurs intonations, leur phrasé, malgré l'anonymisation. Nous avons donc décidé de privilégier les réseaux qui

[7] La première tranche a été récoltée entre 1968 et 1971 (*cf.* Blanc et Biggs, 1971).

[8] Chambre des Notaires de Paris, valorisation de l'indice des Notaires 2005 ; pour Neuilly, les renseignements proviennent du site http://www.linternaute.com/ville/ville/accueil/25124/neuilly-sur-seine.shtml

nous permettaient de trouver des informateurs plus faciles à convaincre. Les enquêtés sont le plus souvent, des personnes auprès de qui des amis nous ont introduits. L'idée de sélectionner quelques arrondissements témoins a laissé place à la recherche de locuteurs témoins sur l'ensemble de Paris.

Au reste, comme le rappelle B. Lahire (1998 : 20-21), lorsqu'un échantillon de locuteurs est limité (dans notre cas à une centaine d'informateurs) et qu'une catégorie sociale s'incarne donc dans quelques individus, il y a toute chance que leur comportement singulier ne coïncide pas avec les propriétés statistiques attribuables à leur catégorie. Ceci vaut d'autant plus que les Parisiens que nous avons interviewés sont porteurs d'identifications complexes (tel titulaire d'un diplôme scolaire très modeste adore la lecture, telle autre a milité et a dû prendre la parole en public, incorporant ainsi une partie des conventions stylistiques propres à la langue politique légitime). Nous avons donc maintenu l'exigence de diversification des locuteurs et de métadonnées, tout en minimisant l'enjeu d'un échantillonnage rigoureusement équilibré.

D'autre part les trentenaires se sont révélés être une catégorie intéressante. Si les adolescents du corpus assument souvent le rôle d'innovateurs, les trentenaires des quartiers populaires ont revendiqué leur rejet des parlures jeunes.

1.5. Du privilège accordé aux natifs à la prise en compte des nouveaux arrivants

L'hétérogénéité linguistique qui résulte des mélanges de populations caractéristiques des centres urbains est l'objet de la sociolinguistique : dès ses premiers travaux, le fondateur de cette discipline, W. Labov (1976), s'est effectivement intéressé à l'étude de la stratification sociale des variantes et à leur impact sur le changement linguistique. Labov laissait cependant de côté le multilinguisme, bien que les villes américaines se caractérisent par un grand mélange de populations, mais les travaux sur le multilinguisme ont démarré assez rapidement. Dans le domaine francophone, L.-J. Calvet s'est penché sur les villes d'Afrique et du Moyen-Orient à partir de 1994[9]. En France, des travaux sont encore menés notamment à Grenoble (C. Trimaille 2004), à Paris (C. Saillard et J. Boutet pour le chinois ; C. Juillard pour les langues d'Afrique), à Rennes (T. Bulot), etc.[10], pour envisager les dynamiques linguistiques que créent les contacts des langues de l'immigration avec la langue dominante.

Paris est un lieu d'immigration traditionnel. La ville accueille des provinciaux depuis des siècles. À partir du début du XX^{e} siècle, sont venus s'ajouter des migrants originaires d'Europe, en particulier d'Italie, d'Espagne, du Portugal et d'Europe de l'Est. Enfin depuis une cinquantaine d'années, avec la mondialisation, les grands flux migratoires concernent aussi le Maghreb, le Moyen Orient, l'Afrique noire, l'Asie. On peut faire l'hypothèse que le français a évolué et continue à le faire au contact des langues d'origine de ces migrants.

Or, dans le projet initial de CFPP2000, les natifs étaient privilégiés : l'objectif de la première phrase de recueil des données était de constituer une base permettant de décrire les variations de la langue « commune », afin de les distinguer éventuellement des variantes qui résultent de l'impact des langues premières des migrants sur leur français qui reste du « français langue seconde ». Il nous avait paru important de ne pas rabattre la situation de la France, pays à idéologie monolingue où l'administration, l'école qui scolarise tous les enfants, les médias, le monde du travail sont massivement francophones, sur la situation africaine où le plurilinguisme n'a pas le même statut. Dans une première étape, seuls ceux qui sont nés ou qui sont arrivés avant l'âge de sept ans dans les zones d'enquête ont donc été pris en compte[11].

[9] Il s'agit toutefois d'observations sur la répartition fonctionnelle des langues et non de collectes de corpus.

[10] Ces chercheurs ont collecté leurs corpus avec un grand souci d'immersion dans les milieux observés. Leurs recherches "écologiques" posent cependant des problèmes de non-comparabilité. Et surtout, elles sont centrées sur les milieux issus de l'immigration, alors que, paradoxalement, les contributions accessibles sur la (ou les) variété(s) des Parisiens ordinaires restent rares ou bien sont centrées sur la variation phonétique (Detey, Durand, Laks et Lyche (éds.), 2010).

[11] Dès la première tranche des locuteurs qui ne sont pas d'origine parisienne apparaissent : ainsi, dans le XVIII^{e}, Pierre-Marie Simo

1.6. L'élargissement du corpus aux nouveaux arrivants

Sans renoncer à cet objectif, il a paru utile de compléter le corpus principal par des études de cas destinées à mettre en évidence les zones du français des migrants qui résistent à l'acculturation linguistique. Pour ces études de cas, nous ferons varier les langues premières selon leur distance au français (langues romanes, indo-européennes, non indo-européennes) ; selon que les locuteurs appartiennent à des couches populaires ou privilégiées ; selon leur sexe ou encore selon que les migrants vivent dans un milieu leur permettant de maintenir l'usage de leurs langues ou que des intermariages ont entraîné l'abandon des langues d'origine.

Pour quelques-uns de ces locuteurs-témoins, nous contrasterons le français « légitime » de l'interview avec le français des conversations entre pairs.

2. Enregistrement, transcription et diffusion

2.1. Enregistrement

Le matériel utilisé, des enregistreurs *firewire Tascam H2P2*, a permis d'obtenir des enregistrements de bonne qualité. Les fichiers audio ont ensuite été anonymés avec *Audacity*[12]. Tous sont disponibles sous deux formats : WAVE et MP3 (version plus légère, mais de qualité sonore moins bonne).

2.2. Transcription

On compte habituellement 40 heures de travail pour une transcription (orthographique) d'une heure d'oral sur du monologue, ce qui peut s'élever à 70 heures et plus pour les dialogues et les multilogues en raison des chevauchements entre locuteurs et des bruits variés plus nombreux.

Le logiciel de transcription utilisé est *Transcriber*[13] qui permet de réaliser des transcriptions de l'oral alignées avec le signal. Les raisons de notre choix tiennent à la facilité d'utilisation de ce logiciel aussi bien pour le travail de transcription que pour la lecture synchronisée de la transcription et du signal audio (même pour des utilisateurs novices). Les fichiers de transcription apparaissent sur le site dans le format original produit par *Transcriber* permettant de coder la synchronisation du son et de sa transcription : ces fichiers sont au format XML et leur nom de fichier a en général une extension du type « .trs ». On trouve aussi en ligne d'autres présentations de ces données de transcription : l'une permet de lire de manière plus aisée le contenu de la transcription, l'autre permet d'accéder directement en ligne à une présentation synchronisée du texte de la transcription et du signal audio associé[14] (sans passer par *Transcriber*). Le corpus a été découpé au niveau du tour de parole[15] dans sa transcription via Transcriber.

La **Figure 1** donne à voir l'alignement du fichier audio et de la transcription pour l'une des interviews de CFPP2000 dans l'interface proposée par *Transcriber*. La fenêtre principale contient la transcription découpée en tours de parole. Les boutons sous la transcription permettent de contrôler l'écoute du signal. Celui-ci est visible dans la partie basse de la figure juste au dessus du redécoupage en énoncés

qui est arrivé à 9 ans du Cameroun ; un des enquêteurs est anglais et n'est venu à Paris qu'une fois son doctorat achevé. D'autre part, des provinciaux sont aussi représentés : dans le XII[e] Valentine Testanier et Thérèse Le Vern viennent l'une de Normandie et l'autre de la Réunion et les deux enquêtrices sont provinciales.

[12] http://audacity.sourceforge.net/

[13] http://trans.sourceforge.net/en/presentation.php

[14] Cette lecture synchronisée est rendue disponible par l'utilisation d'outils développés par le CRDO.

[15] Quelques très longs passages ont été découpés en séquences dans certains corpus.

(par intervenant). La dernière ligne indique le moment courant (en secondes) depuis le début de l'enregistrement.

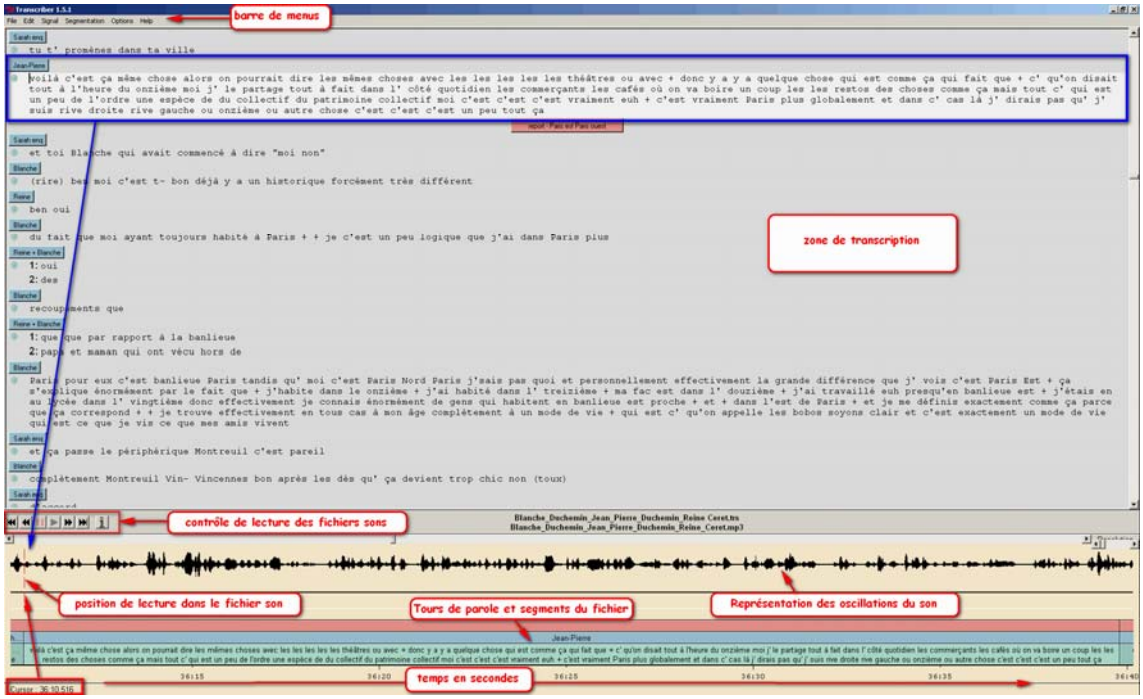


Figure 1 : Interface de Transcriber

La **Figure 2** donne à voir uniquement le fichier de transcription exporté à partir de *Transcriber* (au même « endroit » que dans la figure précédente).

```
<Turn speaker="spk2" startTime="2170.516" endTime="2199.786">
<Sync time="2170.516"/>
voilà c'est ça même chose alors on pourrait dire les mêmes choses avec les les les les les théâtres ou avec + donc y a y a quelque chose qui est comme ça qui fait que + c'
qu'on disait tout à l'heure du onzième moi j' le partage tout à fait dans l' côté quotidien les commerçants les cafés où on va boire un coup les les restos des choses comme
ça mais tout c' qui est un peu de l'ordre une espèce de du collectif du patrimoine collectif moi c'est c'est c'est vraiment euh + c'est vraiment Paris plus globalement et
dans c' cas là j' dirais pas qu' j' suis rive droite rive gauche ou onzième ou autre chose c'est c'est c'est un peu tout ça
</Turn>
</Section>
<Section type="report" topic="to8" startTime="2199.786" endTime="2541.325">
<Turn speaker="spk1" startTime="2199.786" endTime="2202.069">
<Sync time="2199.786"/>
et toi Blanche qui avait commencé à dire "moi non"
</Turn>
<Turn speaker="spk4" startTime="2202.069" endTime="2206.342">
<Sync time="2202.069"/>
(rire) ben moi c'est t- bon déjà y a un historique forcément très différent
</Turn>
<Turn speaker="spk3" startTime="2206.342" endTime="2206.700">
<Sync time="2206.342"/>
ben oui
</Turn>
<Turn speaker="spk4" startTime="2206.700" endTime="2212.950">
<Sync time="2206.700"/>
du fait que moi ayant toujours habité à Paris + + je c'est un peu logique que j'ai dans Paris plus
</Turn>
<Turn speaker="spk3 spk4" startTime="2212.950" endTime="2213.356">
<Sync time="2212.950"/>
<Who nb="1"/>
oui
<Who nb="2"/>
des
</Turn>
<Turn speaker="spk4" startTime="2213.356" endTime="2214.076">
<Sync time="2213.356"/>
recoupements que
</Turn>
<Turn speaker="spk3 spk4" startTime="2214.076" endTime="2215.544">
<Sync time="2214.076"/>
<Who nb="1"/>
que que par rapport à la banlieue
<Who nb="2"/>
papa et maman qui ont vécu hors de
</Turn>
<Turn speaker="spk4" startTime="2215.544" endTime="2255.897">
<Sync time="2215.544"/>
Paris pour eux c'est banlieue Paris tandis qu' moi c'est Paris Nord Paris j'sais pas quoi et personnellement effectivement la grande différence que j' vois c'est Paris Est +
ça s'explique énormément par le fait que + j'habite dans le onzième + j'ai habité dans l' treizième + ma fac est dans l' douzième + j'ai travaillé euh presqu'en banlieue est
+ j'étais en au lycée dans l' vingtième donc effectivement je connais énormément de gens qui habitent en banlieue est proche + et + dans l'est de Paris + et je me définis
exactement comme ça parce que ça correspond + + je trouve effectivement en tous cas à mon âge complètement à un mode de vie + qui est c' qu'on appelle les bobos soyons clair
et c'est exactement un mode de vie qui est ce que je vis ce que mes amis vivent
</Turn>
```

Figure 2 : Extrait de transcription brute

Ce fichier (au format XML) organise dans le temps les tours de parole (marqué ci-dessus par un empilement) par un jeu de jalons textuels (des balises XML) structurant ainsi l'ensemble des informations, il intègre aussi des informations sur les intervenants, les thématiques du discours…

La **Figure 3** décrit de manière schématique l'organisation structurelle de ce fichier :

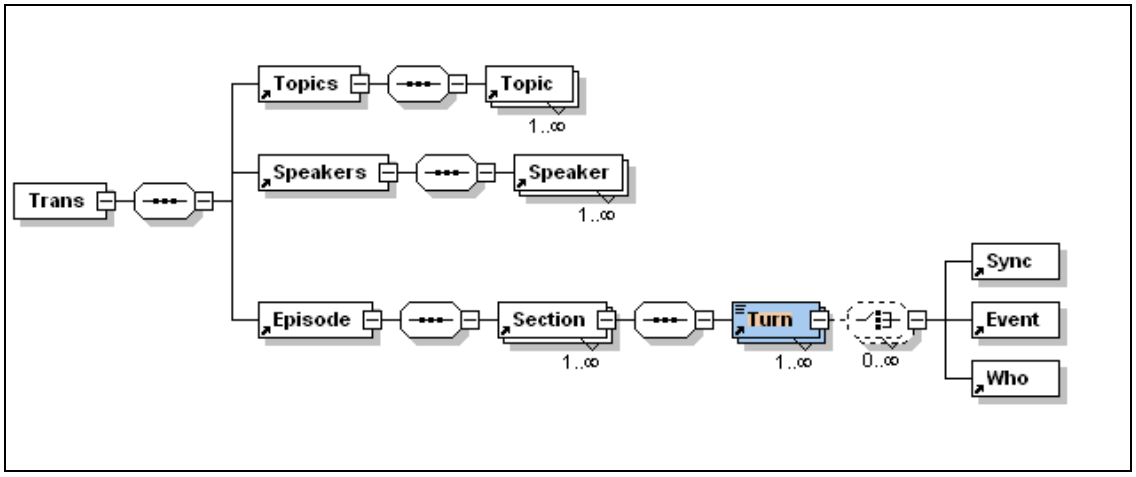


Figure 3 : Arborescence d'un fichier de transcription

Ce schéma décrit de manière plus générale la structure des informations dans un fichier de transcription pour *Transcriber*. Le nœud *Trans* est la racine du document ; on y trouve ensuite des informations générales sur les thèmes, sur les intervenants (nœuds *Topics* et *Speakers*). *Transcriber* permet de segmenter le signal en sections, en tours de parole (dans les éléments associés aux nœuds de la branche inférieure de l'arbre ci-dessus).

Par ailleurs, on peut, dans l'interface de *Transcriber*, faire des recherches de chaîne de caractères dans une transcription, ce qui permet de retrouver facilement n'importe quel passage.

Un mode d'emploi[16] de *Transcriber* est disponible sur le site CFPP2000.

3. Conventions de transcription

Toute transcription est un compromis forcément boiteux entre le respect des particularités orales et la lisibilité. Afin de faciliter la lecture[17] et de simplifier les traitements automatisés du corpus, nous avons adopté le code orthographique avec quelques aménagements, suivant en cela les pratiques initiées par les chercheurs québécois, par ESLO, par le DELIC et VALIBEL[18]. L'alignement sur le son permettant de toute façon de revenir facilement à la version orale, nous avons décidé de ne pas essayer de rendre compte de l'intonation ou des variantes phonétiques.

3.1. Identification des corpus et des locuteurs

Chaque fichier est identifié au moyen d'un code individuel, par exemple :
`[CFPP2000 03-01] Ozgur_Kilic_H_32_alii_3e` ; la partie entre crochets est l'identifiant du fichier du corpus CFPP2000, suivi du numéro de l'arrondissement ou des initiales de la banlieue où a été réalisé

[16] http://cfpp2000.univ-paris3.fr/tools/Utilisation_de_transcriber.ppt
[17] Il s'agit de ne pas rebuter des utilisateurs non linguistes, mais aussi de permettre une lecture cursive à des linguistes. On sait que même des spécialistes ont une lecture extrêmement ralentie lorsque les conventions s'écartent trop de leurs habitudes.
[18] http://sites.univ-provence.fr/delic/corpus/conventions.html; http://www.uclouvain.be/81836.html

l'enregistrement et d'un chiffre correspondant à l'ordre d'enregistrement dans la base de données ; viennent ensuite le pseudonyme du locuteur principal, son sexe (F ou M), son âge, et le numéro de l'arrondissement. Le pseudonyme est choisi autant que faire se peut pour évoquer les caractéristiques du nom de départ (origine géographique, notoriété générationnelle des prénoms, etc). Dans l'identifiant `CFPP2000 [03-01] Ozgur_Kilic_H_32_alii_3e`, le locuteur principal est, par convention Özgur Kilic et ce pseudonyme évoque son origine turque ; *alii* signale que les autres participants trop nombreux n'ont pas été énumérés. Ils reçoivent aussi des pseudonymes qui permettent de les identifier et qui sont récapitulés dans la fiche descriptive, par exemple Michel Chevrier participe au même enregistrement.

3.2. Une transcription orthographique avec quelques aménagements

Nous transcrivons les mots en orthographe sans correction des écarts à la norme, dès lors que le segment correspond à un morphème attesté en français. Ainsi, bien que ce soit une femme qui s'exprime, nous écrivons *mis* et non *mise* lorsqu'à l'oral la locutrice n'a pas réalisé l'accord du participe :

```
c'est pas pour des raisons euh catholiques qu'ils m'ont mis dans une école de
bonnes sœurs
CFPP2000 [11-01] Anita Musso-F-46-11e
```

En effet, *mis* est un morphème du français. Selon le même principe, nous pouvons « respecter » un futur comme « cousera » (au lieu de *coudra*) qui se laisse ramener à une succession de morphèmes français : radical du verbe + *-er-*, morphème du futur, succession présente pour de nombreux verbes + morphème de troisième personne, *-a*.

En revanche, les variantes de prononciation qui n'ont pas de correspondance orthographique reçue dans les dictionnaires ne sont pas notées : nous n'écrivons pas « f'nêt », « f'nêtr », pour *fenêtre*, ce qui rendrait très difficile les recherches assistées par l'ordinateur dans les corpus. Nous pouvons cependant noter la variation entre « oui » et « ouais », puisque « ouais » est un morphème recensé par les dictionnaires.

Nous ne notons pas non plus les allongements de syllabes, ou les *e* qui apparaissent en fin de mot (la frontière avec des « euh » d'hésitation ou de remplissage est incertaine).

Nous avons cependant – et c'est un point de divergence avec d'autres projets – élidé les clitiques quand cette élision s'entendait à l'oral. Nous écrivons alors : *j'* devant consonne, *t'* devant voyelle *: j'sais ; t'arrives.* Ces graphies se sont bien répandues, gagnant peu à peu la bande dessinée, le roman, les blogs. Les adopter amène à produire une description plus proche de la réalité morphologique. Il est faux par exemple que la forme sujet et la forme objet direct du pronom de deuxième personne soient en distribution complémentaire, puisque *tu* et *te* sont parfois neutralisés sous la forme *t'*. Cette décision ne devrait pas poser de problèmes puisque les morphèmes « t', j', qu', c', s', d', n' » existent et que la liste en est fermée :

```
Lies : en fait tu t'rends compte que finalement tous les vendeurs sont parlent
arabe et donc en plus avec leur client t'as l'impresssion qu- qu'ils sont
vachement potes et toi t'arrives bon j'suis sûre qu'on paie plus que les autres
[CFPP2000-18-01] Paul Simo (…) Lies Simo, F-30 néerlandaise
```

En revanche nous nous refusons à « i » ou « iz » pour *ils* qui s'écartent des habitudes orthographiques du français.

Nous avons respecté les apocopes qui constituent selon nous des néologismes (cela conduira à revoir les règles de lemmatisation des concordanciers et à poser des équivalences entre : « appart- ; prof ; aspi-, etc. ; » et « appartement, professeur, aspirateur, etc. »).

Les emprunts sont autant que possible transcrits selon l'orthographe usuelle dans leur langue d'origine. Lorsque le mot est inconnu, ou qu'il n'a pas été répertorié par les dictionnaires parce qu'il est non standard, il est transcrit selon une graphie qui se rapproche des séquences usuelles du français. Pour les mots dits « des banlieues », nous avons eu recours à des dictionnaires comme Le *Dictionnaire de la Zone*[19].

Nous employons des majuscules à l'initiale des noms propres ; cependant nous avons renoncé à les isoler par des signes spéciaux, leur délimitation posant des problèmes qui nécessitent une étude spéciale.

Les passages incompréhensibles sont représentés par un (pour une syllabe) ou plusieurs X majuscules. Par convention, comme le font Delic et Valibel, nous nous limitons à 3 X.

Nous suivrons les conventions de Valibel pour opposer les sigles transcrits en capitales sans points : *SNCF* et les acronymes transcrits par une majuscule au début du mot, le reste restant en bas de casse comme un nom propre ordinaire : *Fnac*.

3.3. Amorce, multi-transcription et alternances orthographiques

L'amorce d'un mot est notée par un tiret accolé au mot :

```
un mi-
```

Pour faciliter la lecture, nous pensons nécessaire de restreindre la notation des alternances de transcription. Nous nous contentons de noter l'interprétation la plus probable. Lorsqu'on ne peut éviter de poser une multi transcription, nous suivons les conventions de DELIC en mettant entre parenthèses les deux possibilités séparées par une virgule :

```
(d'accord, d'abord).
```

Les alternances orthographiques n'ont pas été notées systématiquement. Lorsqu'elles le sont, c'est entre-parenthèses, conformément à l'usage du DELIC :

```
on (n') a pas
```

3.4. Liaisons

Les liaisons non standard ont été indiquées par un « z » ou un « t » entre tirets. Ce principe s'étend à des cas où le statut du « z » n'est pas très clair et où il pourrait s'apparenter à un morphème flottant ou à une variante morphologique du morphème suivant :

```
Mille-z-assiettes
donne moi-z-en
des chefs de gare-z-en retraite
```

3.5. Pauses et ponctuation

Nous n'utilisons ni les points ni les virgules. Il aurait peut-être été intéressant d'indiquer par une ponctuation forte ou demi-forte toutes les segmentations en unités de discours que le transcripteur perçoit, – que son impression soit due à une pause, par un allongement vocalique ou par un contour intonatif descendant. C. Blanche Benveniste (2007) est d'ailleurs revenue sur la proscription qu'elle avait

[19] http://www.dictionnairedelazone.fr/

largement contribué à établir. Nous avons finalement maintenu pour cette première tranche les conventions suivantes :

+	pause brève ;
++	pause longue ;
///	interruption du discours.

Le point d'interrogation, le point d'exclamation et les guillemets ont été utilisés lorsque le transcripteur entendait nettement l'intonation :

?	interrogations avec montée de la voix
!	Exclamation
« »	Les décrochages liés au discours direct sont signalés. Les transcripteurs ont décidé de ne pas noter la frontière droite lorsqu'ils hésitaient.

3.6. Chevauchements et tours de parole

Le système *Transcriber* est mal adapté à la notation des chevauchements, en particulier il ne permet pas de noter le cas où plus de deux personnes prennent la parole simultanément. Ces chevauchements multiples sont rares dans notre base. Ils ne sont pas pris en compte. Par ailleurs, nous avons décidé pour des raisons d'exploitation informatique de ne pas couper les mots, même lorsqu'une partie seulement est concernée par le chevauchement. C'est pourquoi, nous n'avons pas ajouté de symboles permettant de délimiter le début et la fin des paroles prononcées. Ces indications resteraient de toute façon approximatives. Un retour au son s'impose donc si l'on veut travailler sur les interactions.

Nous n'avons pas interrompu le tour par un retour à la ligne, lorsque le locuteur principal poursuit sa prise de parole et qu'un second locuteur intervient en arrière-plan en se bornant à des « mm » approbatifs ou à des interjections (hum) sans interrompre le tour : nous faisons figurer entre parenthèses ces régulateurs. Ces interventions n'ont pas été notées dans les transcriptions réalisées jusqu'en avril 2009. (Les « mm » du locuteur principal ne sont pas mis entre parenthèses).

Nous notons également entre crochets des bruits tels que les rires ou la toux. Dans *Transcriber*, ils apparaissent sous cette forme :

```
j'l'ai fait non j'ai pas arrêté ils m'ont renvoyé j'vais vous dire comment ils ont
fait [rire]
```

Dans la version en ligne via *Real Player*, ils apparaissent avec le codage qui permet leur non prise en compte par les outils informatiques, concordanciers, étiqueteurs, etc. :

```
j'l'ai fait non j'ai pas arrêté ils m'ont renvoyé j'vais vous dire comment ils ont
fait[rire|noise|instantaneous]
[CFPP2000-SO-02] Youcef-Zerari, 29 ans
```

4. Accessibilité (problèmes juridiques) et diffusion

Nous avons demandé leur consentement aux locuteurs que nous avons interviewés. Nous leur avons laissé un document mentionnant les finalités de l'enquête et fournissant le nom et l'adresse du responsable de l'enquête. Les interviewés ont accepté par écrit de nous confier leurs données en sachant que celles-ci devaient être transcrites et mises en ligne. Il a été impossible de leur signaler les types d'utilisation de ces données puisque les corpus sont conçus pour des usages pluriels et que dès

lors nous ne sommes pas en mesure d'énumérer les types d'exploitation futurs. Nous avons précisé aux enquêtés que les noms propres des personnes seraient anonymisés, sauf en cas de demande expresse de leur part. Nous leur avons également signalé qu'ils pouvaient nous demander de couper certaines parties de l'interview (ce qui a d'ailleurs été parfois le cas).

Les enseignants qui sont à l'origine de CFPP2000 ont estimé que, rémunérés par l'université en tant qu'enseignants-chercheurs, ils pouvaient mettre à disposition de leurs collègues les ressources qu'ils collectaient. En attendant la création des corpus oraux informatisés de plusieurs millions de mots indispensables aux études quantitatives[20], la mutualisation des données reste une solution, quels que soient les problèmes d'harmonisation qu'elle pose. C'est pourquoi les données de CFPP2000 sont libres d'accès, disponibles sous contrat Creative Commons[21]. En ne réservant pas l'accès du corpus aux seuls chercheurs, nous voulons aussi rendre aux enquêtés ces paroles qu'ils nous ont confiées. La restitution des données fonctionne comme une sorte de contrôle *a posteriori*. Au-delà, comme l'écrit Louis Quéré (2002 : 81), les sciences sociales « se trouvent devoir rendre des comptes, […] à un public très large, celui de tous ceux qui sont susceptibles d'être concernés par les problèmes de leur société à un moment donné ».

5. Les métadonnées du corpus

L'ensemble des données du corpus a été associé à des métadonnées permettant de décrire ces données : l'idée sous jacente étant à la fois de décrire les données du corpus (dans une perspective d'archivage de grands corpus où il est essentiel de pouvoir commodément retrouver un enregistrement) mais aussi de pouvoir travailler en sélectionnant des données correspondant à des caractéristiques particulières (rechercher par exemple les entretiens correspondant à des participants ayant un profil donné (homme, femme, âge, profession etc.).

Le processus de description des données aboutissant à une première phase de métadonnées s'est déroulé en 2 temps :

- Chaque entretien a donné lieu à une fiche non normalisée décrivant l'entretien et ses participants[22]. On trouve sur le site le formulaire (au format PDF) produit pour chaque entretien. Voici par exemple la description d'une enquêtée :

```
Locuteur principal
Rôle ; interviewé
État civil
Prénom et nom : Mira MARKOVIC
Sexe : Féminin
Lieu de naissance, Belgrade
Habitat pendant la petite enfance Serbie
Mobilité géographique : Arrivée à Paris à 28 ans
Relation avec l'enquêteur : lien d'amitié
Réseau par lequel l'enquêté a été contacté
Âge, date de naissance : 88 ans ; née le 02.05. 1922
Situation familiale : mariée,
scolarité : Supérieur (bac + trois ans d'école des Beaux-arts)
Parcours professionnel
    A exercé les activités de : artiste peintre
Parents : père mathématicien ; mère femme au foyer
Langues parlées par le locuteur : serbe, français
```

[20] En 2004, le British National Corpus comportait 10 millions d'occurrences transcrites de l'anglais parlé.

[21] http://creativecommons.org/

[22] On retrouve *infra* une trace de ce formulaire dans le descripteur *dc:description* utilisé dans le langage de métadonnées Dublin Core

- L'ensemble de ces formulaires a ensuite été utilisé pour produire des métadonnées en respectant les recommandations OLAC/Dublin Core.

La communauté OLAC (*Open Language Archive Community*) est l'émanation d'une collaboration entre trois organismes linguistiques internationaux : le LDC (*Linguistic Data Consortium*), le SIL (*Summer Institute of Linguistics*) International et la liste de diffusion LINGUIST. Les métadonnées proposées dans ce projet sont une spécialisation des métadonnées du Dublin Core. Le projet OLAC offre ainsi un jeu minimal de méta-données et une interface permettant de référencer la description de ressources linguistiques. Les spécifications proposées ne remplacent pas les métadonnées du Dublin Core, mais les spécifient par rapport aux attentes de la communauté linguistique. Cinq extensions au Dublin Core concernant la description de ressources linguistiques ont été proposées :

- Type de discours (Discourse Type) : comprend des types du genre « dramatique », « narration », « jeux de mots », etc.
- Identification de la langue (Language Identification) : fait référence aux codes ISO (« fr » pour français, « en » pour anglais, etc)
- Champ linguistique (Linguistic Field) : propose des champs tels que sociolinguistique, phonétique, etc.
- Type de données linguistiques (Linguistic Data Types) : on retrouve ici trois types de données, « lexique », « texte-primaire » et « description de langue ».
- Rôles des participants (Participant Roles) : liste des différents rôles que peuvent avoir les participants, tels qu'annotateur, auteur, locuteur, etc.

Pour CFPP2000, nous avons initié le travail de construction des métadonnées des données du corpus en nous appuyant sur OLAC, nous avons ensuite surchargé (en dehors des préconisations d'OLAC) le descripteur *dc:description* avec un contenu structuré de la manière suivante :

```
Entretien -> [
    Langue -> xx
    Editeur -> xx
    Date -> xx
    Lieu -> xx
    Enquêteur :-> xx
    Quartier concerné -> xx
    Transcription -> xx
    Anonymisation -> xx
]
Enquêté X -> [
    Etat civil -> [
        Prénom et NOM fictif -> xx
        Sexe -> xx
        Age au moment de l'enregistrement -> xx
        Situation familiale : mariée, deux enfants
        Relation (parenté, ami, etc) avec Enquêté 2 -> xx
        Relation (parenté, ami, etc) avec Enquêté 3 -> xx
    ]
    Réseau par lequel contacté -> xx
    Scolarité -> [
        Dernier diplôme obtenu -> xx
    ]
    Commentaires -> xx
    Travail -> [
        Activité actuelle -> xx
        Activités passées -> xx
    ]
     Parents (profession, lieu de naissance, scolarité…) -> [
        Mère -> xx
        Père -> xx
    ]
]
```

Le descripteur *dc:description* intègre ainsi de manière structurée des informations difficilement intégrables dans les descripteurs proposés par la norme OLAC. Il reprend en partie le contenu de la fiche descriptive initiale de chaque entretien.

Cette extension du modèle fourni par OLAC n'est pas une solution définitive mais participe d'une réflexion en cours sur la mise au point de métadonnées pour les données orales, cette réflexion menée avec d'autres chercheurs (en particulier les participants au projet Rhapsodie) vise à proposer des extensions aux normes actuellement disponibles pour les enrichir en tenant compte des faiblesses actuelles de ces modèles par rapport aux besoins des chercheurs pour décrire au mieux leurs données.

Chaque entretien du corpus CFPP2000 est donc associé in fine à une fiche de métadonnées suivant les recommandations OLAC / Dublin Core (le descripteur *description* proposé par cette norme, servant provisoirement d' « abri » permettant d'intégrer toutes les informations que nous avons jugé nécessaires et qui n'étaient pas prises en charge par OLAC/Dublin Core).

Chaque fiche de métadonnées est un fichier au format XML disponible en ligne ; pour faciliter sa lecture, cette fiche est accompagnée d'une feuille de styles permettant de lire son contenu de manière aisée dans un navigateur web.

Le catalogue constitué par l'ensemble des métadonnées des entretiens de CFPP2000 est associé à un moteur de recherche permettant de sélectionner des descripteurs contenant une information donnée puis de retourner aux entretiens correspondants à ces valeurs de descripteurs. On peut ainsi sélectionner des entretiens sur la base des métadescriptions établies en amont pour les décrire.

Le travail amorcé ici pour construire des métadonnées pertinentes pour décrire au mieux des données orales doit s'accompagner d'un enrichissement des modèles actuellement disponibles pour décrire les données de langue. Ce développement (déjà accompli dans d'autres domaines de connaissance) est un chantier incontournable pour la diffusion de données orales et leur exploitation future par des linguistes.

6. Premiers résultats

6.1. L'hypothèse du français parisien commun

Le français pratiqué est un français que nous avons appelé « français commun parisien ». Ce singulier ne va pas de soi. À Paris et dans la banlieue autour de Paris, la diversité des origines des locuteurs de première et deuxième génération entraîne la co-présence de nombreuses langues. Nous faisons cependant l'hypothèse qu'à l'exception des émigrés de première génération, les Parisiens natifs (ou arrivés avant sept ans) usent au cours de leurs échanges avec les enquêteurs d'un « français de communication » 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 registre oral adopté par la dyade enquêteur/enquêté(s) et qui est considéré comme approprié pour une conversation entre des gens qui n'appartiennent pas au même espace intime ou au même réseau amical ou professionnel s'écarte sur plusieurs aspects du français *standardisé* des grammaires :

Le parisien "commun" que nous recueillons est de l'oral, alors que les grammaires décrivent surtout de l'écrit normé. Les énoncés font largement appel aux structures à présentatifs ou aux structures clivées. Par ailleurs, les hésitations, inachèvements, reformulations entraînent des discontinuités dans

leur structure. Ces caractéristiques bien répertoriées par les spécialistes depuis les travaux de C. Blanche-Benveniste contrastent avec les modes d'organisation spécifiques de l'écrit normé.

Les grammaires tardent à intégrer des évolutions pourtant bien avancées qu'il s'agisse de prononciations (avec, par exemple, le recul des liaisons) ou de constructions syntaxiques (comme les interrogatives indirectes en *qu'est-ce que*) .

La pression normative paraît reculer dans certains domaines. Les enquêtes montrent la faible observance de multiples et minuscules interdits ou recommandations qui se transmettaient parfois depuis des siècles dans des recueils de fautes, et qui ont alimenté bien des passions. (On peut citer l'usage largement répandu de *ouais* que le TLF considère encore comme familier voire dans certains emplois franchement vulgaire).

6.2. Quelques traces du vernaculaire

Quelques termes considérés comme populaires ou familiers par les dictionnaires se retrouvent dans ces enquêtes, le plus souvent quand les locuteurs sont jeunes, parfois également lorsqu'ils sont plus âgés. CFPP2000 permet de contextualiser ces formes pour voir comment elles sont employées : les locuteurs à fort capital scolaire les emploient-ils comme des marqueurs discursifs, comme des ressources permettant de produire des effets de sens particuliers (humour, connivence, goût du décalage, etc.) ?

6.3. Les discours produits « sur la ville » et les choix de registres de langue.

À travers la dynamique de l'interview, les locuteurs construisent des référents spatiaux, et des identités urbaines conflictuelles qui sous-tendent leur façon d'habiter la ville. Selon qu'ils se réfèrent à Paris, ou qu'ils ont l'impression de vivre dans un archipel de communautés divisées, qu'ils se posent comme membre de la collectivité urbaine, ou qu'ils disent appartenir à des groupes dissidents, leur registre de langue est différent.

Nous nous intéresserons également aux formes en mention imputées à d'autres locuteurs-énonciateurs quand ils sont, en outre, envisagés dans leur altérité (épinglés par exemple en tant que « jeunes », par des locuteurs âgés ou que « vieux » par des locuteurs jeunes ; ou bien). Il s'agit le plus souvent de vocabulaire comme dans cet exemple ou une trentenaire évoque les façons de parler de son jeune frère :

quand j' lui présente une fille qui est un peu qui a mon âge il dit "oh on dirait une daronne ! (cfpp2000.[11-04] Amélie Tourette-28 ans)

7. Travaux effectués à partir de CFPP2000

Barbéris, J.-M., 2010a, « 'Quand t'es super bobo'… La deuxième personne générique dans le français parisien des jeunes », Actes du *CMLF2010,* publié en ligne (http://dx.doi.org/10.1051/cmlf/2010258)

Branca-Rosoff, S., (à paraître) « « Les variations langagières dans le lexique du corpus du français parlé parisien (CFPP2000). Un outil pour le FLE ? », *Hommage au professeur Marazza*, coordonné par le Professeur M. Margarito de l'U. de Turin.

Branca-Rosoff, S., (à paraître) « 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 », *Cahiers de praxématique*.

Branca-Rosoff, S., et Fleury, S., 2011, « Informatique et linguistique. Dialogues autour du couple Futur simple *versus* Futur périphastique », colloque de l'ASL, "Sciences du langage et Nouvelles technologies", Lambert Lucas.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., M. Pires, 2011, « Constitution et exploitation d'un corpus de français parlé parisien. Contraintes et apports possibles de la langue au texte », *Corpus* 10 "varia", 81-98 (et en ligne "varia" **http://corpus.revues.org/**)

Détrie, C., 2010, « De *voir* à *tu vois* / *vous voyez* : fonction sémantico-énonciative et postures énonciatives construites par ces particules interpersonnelles », Congrès Mondial de Linguistique Française 2010, Neveu F., Muni Toke V., Klingler T., Durand J., Mondada L. et Prévost S. (éd.), texte en ligne <http://www.linguistiquefrancaise.org/> et CD-Rom, 755-766.

Fleury, S., et 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, *cahiers AFLS* 16, 1 (2010) 63-98

Lefeuvre F. & Moline E. éds, 2011, *Unités syntaxiques et unités prosodiques*, *Langue française*, n° 170.

Lefeuvre F. & Moline E., 2011, « Présentation : Unités syntaxiques et unités prosodiques », *Langue française*, 170, *Unités syntaxiques et unités prosodiques* (Lefeuvre F. & Moline E. éds), 3-10.

Lefeuvre, F., et Moline, E., 2011, « Unités syntaxiques et unités prosodiques : bilan des recherches actuelles », Langue française, 170, Unités syntaxiques et unités prosodiques (Lefeuvre, F., et Moline, E., éds), 143-157.

Lefeuvre, F., 2011, « *Bon* dans le discours oral : une unité averbale autonome ? », *Les énoncés averbaux autonomes entre grammaire et discours*, Ophrys (Lefeuvre et Behr éds), 165-185.

Lefeuvre, F., 2011, (sous presse) « *Bon* et *quoi* à l'oral : marqueurs d'ouverture et de fermeture d'unités syntaxiques en discours », *Linx* (Krazem ed.)

Lefeuvre, F. (sous presse) « *Bon* » à l'oral en tant que préfixe : étude topologique » (Richard éd.).

Lefeuvre, F., et Morel, M.-A., Teston-Bonnard, 2011 : « Valeurs prototypiques de *quoi* à travers ses usages en français oral », *Neuphilologische Mitteilungen (Bulletin de la Société Néophilologique*, Helsinki), 37-59

Lefeuvre, F. (soumis) « *Eh bien* » comme évaluateur de discours à l'oral (spontané ou représenté), *Travaux de linguistique* (Moline ed.).

Lefeuvre, F. & Tanguy, N., 2011 sous presse, « La représentation de l'oral dans la littérature du 20e siècle : les structures averbales ». In *Quand les genres de discours provoquent la grammaire... et réciproquement,* Despierre C. & Krazem M. (éds), Limoges, Éditions Lambert-Lucas.

Tanguy, N., 2012 à par., « J'ai terminé ma phrase. Ou pas ? L'exemple des compléments différés à l'oral. » In *Ellipses & fragments : morceaux choisis*, Hadermann P., Pierrard M. Roig A. & van Raemdonck D. (éds), Bruxelles, Peter Lang.

Tanguy, N., 2012 à par, « Complémentations en direct. Le fonctionnement des compléments différés à l'oral ». In *Complémentations*, Gautier A., Pino Serrano, L., Valcarcel Ribeiro C. & Van Raemdonck, D., Bruxelles, Peter Lang.

Tanguy N. & Sarda L., 2012 à par., « Comparaison écrit/ oral de *au fond* en français moderne ». *Corpora and Language in Use / Actes du Colloque LPTS 2011 Variation(s) sur la structure de l'oral et l'écrit*, Presses Universitaires de Louvain.

Tanguy, N., 2011 sous presse. « De la prosodie à la syntaxe. Vers une description syntaxique des périodes intonatives ». *Actes de RSL5. Représentations du sens linguistique*, Éditions de l'Université de Savoie.

8. Références bibliographiques

Atlas des franciliens, t. 3 Population et modes de vie, Paris, INSEE.

Baude, O. (coord.), 2006, *Corpus oraux, Guide des bonnes pratiques,* Paris, CNRS éditions et P.U.O.

Bilger, M., Blasco, M., Cappeau, P., Pallaud, B., Sabio, F., Savelli, M.J., 1997, « Transcription de l'oral et interprétation : illustration de quelques difficultés ». *Recherches sur le français parlé*, 14. 57-86.

Bilger, M. (dir.), 2000, « Linguistique sur corpus, études et réflexions », *Cahiers de l'université de Perpignan*, Perpignan, Presses universitaires.

Bilger, M. (éd.), 2000, Corpus, Méthodologie et applications linguistiques, Paris, Champion.

Bilger, M. et Cappeau, P., s.d., « Ce que les corpus nous apprennent sur la langue » <http://icar.univ-lyon2.fr/ecole_thematique/contaci/documents/bilger_cappeau/Bilger-Cappeau-corpus.pdf>

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

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

Blanche-Benveniste, C., (éd.), 1990, *Le Français Parlé: Etudes Grammaticales*, Paris, CNRS.

Blanche-Benveniste, C., 1997, « Transcription et technologie », *Recherches sur le Français Parlé* 14 : 87-100.

Branca-Rosoff, S., à paraître, « 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 », *Cahiers de praxématique*.

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

Branca-Rosoff, S. et Leimdorfer, F. (dirs), 2001, *Langage et Société* 96, « Espaces urbains : analyses lexicales et discursives ».

Branca-Rosoff, S., Fleury, S., Lefeuvre, F. et Pires, M., 2009a, Corpus de français parlé des années 2000 (CFPP2000) <http://cfpp2000.univ-paris3.fr/>

Branca-Rosoff, S., Fleury S., Lefeuvre F., Pires, M. 2009. *Constitution et exploitation d'un corpus de français parlé parisien* <http://cfpp2000.univ-paris3.fr/Presentation.html>

Branca-Rosoff, Grinshpun Y., Régent-Susini A. (éds), *Langue commune* et *changements de normes*, Paris, Honoré Champion.

Bulot, T. (dir.), 2004a, Lieux de ville et identité (Perspectives en sociolinguistique urbaine), volume 1, Paris, L'Harmattan.

Bulot, T. (dir.), 2004b, Lieux de ville et territoires (Perspectives en sociolinguistique urbaine), volume 2, Paris, L'Harmattan.

Bulot, T. et Dubois, L., 2005, *Revue de l'Université de Moncton*, volume 36, numéro 1, 2005

Bulot, T. et Messaoudi, L. (dirs), 2003. Sociolinguistique urbaine, frontières et

Bulot, T. et Veschambre, V., 2004, « Sociolinguistique urbaine et géographie sociale », colloque international Espaces et société aujourd'hui (La géographie sociale dans les sciences sociales et dans l'action) « Sociolinguistique urbaine et géographie sociale : hétérogénéité des langues et des espaces » (Rennes, les 21 et 22 octobre 2004).

Calvet, L.-J., 1994. *Les voix de la ville, introduction à la sociolinguistique urbaine*, Paris, Payot.

http://sites.univ-provence.fr/delic/corpus/conventions.html; http://www.uclouvain.be/81836.html

Durand, J. et Lyche, C., 2004, « Structure et variation dans quelques systèmes vocaliques du français : l'enquête phonologie du français contemporain », *in* Coveney *et al.* (éds.), 217-240.

Detey, S, Durand, J., Laks B. et Lyche Ch (eds.), 2010 : *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement* , Paris, Ed. Ophrys.

Elicop. <http://bach.arts.kuleuven.be/elicop/>

French Language Learning Oral Corpora. <http://www.flloc.soton.ac.uk/>

Fuchs, C. et Habert, B. (éds), 2004, « Le traitement automatique des langues : des modèles aux ressources », *Le français moderne* 72.

Gadet, F., 1997, *Le français ordinaire*, Paris: Armand Colin/Masson, 2ème édition.

Gadet, F., 2003, *La variation sociale en français*, Paris, Ophrys.

Goldman, J.-Ph., 2008, "EasyAlign: a semi-automatic phonetic alignment tool under Praat", <http://latlcui.unige.ch/phonetique/easyalign>

Kerswill, P. & Cheshire J., sd., *Linguistic Innovators: The English of Adolescents in London* <http://www.lancs.ac.uk/fss/projects/linguistics/innovators/overvie>

Kerswill, P. , 2002, "Models of linguistic change and diffusion: New evidence from dialect levelling in British English". Reading Working Papers in Linguistics 6: 187-216.

Kerswill, P. , 2003, "Dialect levelling and geographical diffusion in British English". In D. Britain and J. Cheshire (eds.). Social dialectology. In honour of Peter Trudgill. Amsterdam: Benjamins. 223-243.

Kerswill, P. and Williams, A., 2005, "New towns and koineisation: linguistic and social correlates". Linguistics 43: 1023-1048.

Labov, W., 1976, *Sociolinguistique*, Paris, Minuit, coll. Le sens commun.

Lahire, B., 2001,, *L'Homme pluriel. Les ressorts de l'action*, Paris, Hachette, coll. "Pluriel".

Juillard, C., 1995, *Sociolinguistique urbaine. La vie des langues à Ziguinchor (Sénégal)*, Paris, CNRS éditions.

Lodge, R. A, 2004, *A sociolinguistic history of Parisian French*. Cambridge : CUP.

Milroy, L. et Milroy, J., 1992, « Social networks and social class: Toward an integrated sociolinguistic model », *Language in Society* 21, 1-26.

Mondada, L., 1999a, « L'organisation séquentielle des ressources linguistiques dans l'élaboration collective des descriptions », *Langage et Société*, 89, 9-36.

Mondada, L., 2000, *Décrire la ville. La construction des savoirs urbains dans l'interaction et dans les textes*, Paris, Anthropos.

Quéré, L., 2002, « Pour un calme examen des faits de société », *In* Lahire B., *A quoi sert la sociologie*, Paris, La Découverte, pp. 79-94.

Renaud, P., 2004, « Paroles *dans* et paroles *sur* les pratiques ? Pratiques langagières et sociolinguistique » In I. Léglise (éd.) *Pratiques, langues et discours dans le travail social* , Paris, L'Harmattan.

Saillard C. et Boutet J., 2008, « Construction des répertoires langagiers dans la migration Wenzhou (Chine) à Paris », *Migration et plurilinguisme en France, Cahiers de l'Observatoire des pratiques linguistiques* 2 :72-76.

Sinclair, J., 2001, « Review of Biber D., Johansson S., Leech G., Conrad, S. et Finnegan, E. (1999), *Longman Grammar of Spoken and Written English*, Harlow, Longman », *International Journal of Corpus Linguistics 6*, 339-359.

Stenström, A. B., Andersen, G., Hasund, K., Monstadt, K. et Aas, H., s.d., *User's manual, to accompany The Bergen Corpus of London Teenage Language, COLT)*. Bergen : Department of English, University of Bergen,
<http://torvald.aksis.uib.no/colt/>,
<http://www.hf.uib.no/i/Engelsk/colt/COLTinfo.html>

Thibault, P., 2001, « Regard rétrospectif sur la sociolinguistique québécoise et canadienne », *Revue québécoise de linguistique*, 30/1: 19-42.

Thibault, P. et Vincent, D., 1990, *Un corpus de français parlé*, Québec, CIRAL, collection Recherches sociolinguistiques, 145 p.

Torgersen, E., Kerswill, P. and Fox, S. (2006) Ethnicity as a source of changes in the London vowel system. In F. Hinskens (ed.). Language Variation - European Perspectives. Selected Papers from the Third International Conference on Language Variation in Europe (ICLaVE3), Amsterdam, June 2005. Amsterdam: Benjamins. 249-263.
<http://www.lancs.ac.uk/fss/projects/linguistics/innovators/documents/Iclave3_article_TorgersenKerswillFox_Final_000.pdf>