Pratiques langagières des jeunes de banlieue parisienne : problématique de la constitution d'un corpus

Auphélie Ferreira

Auphélie FERREIRA

Pratiques langagières des jeunes de banlieue parisienne : problématique de la constitution d'un corpus

Abstract

This article deals with how to collect spoken data for studying the language practices of young people who lived in the suburbs of Paris. It is divided into five sections which represents the five steps to realize an oral corpus. This paper will broach the difficulties inherent to the collection of ordinary discourse. Among those difficulties it will tackle the effect of the researcher's interventions on the data collection and how it may influence the analyses based of those data.

Introduction

L'étude des pratiques langagières orales des jeunes de banlieue parisienne nécessite, outre les connaissances théoriques, un soin particulier dans la constitution du corpus. Cette constitution implique différents processus qui peuvent être regroupés sous forme de cinq étapes successives et complémentaires**.** Nous illustrons ces processus à travers la description d'un travail mené auparavant et qui a abouti à un corpus composé du MPF[1] (*Multicultural Paris French*), d'enregistrements récoltés personnellement (dont une partie a rejoint le MPF) et du CFPP2000[2] (Corpus de Français Parlé des années 2000). Cet article dans son ensemble sera l'occasion de montrer que les décisions prises par le chercheur[3] lors de la réalisation d'un corpus reflètent ses objectifs finaux et ainsi qu'il ne semble pas exister de données véritablement « brutes ».

La dénomination « pratiques langagières des jeunes de banlieue parisienne » requiert quelques remarques. La notion de « pratiques langagières » est présentée par Josiane Boutet, Pierre Fiala et Jenny Simonin-Grumbach en 1976. Bien que d'autres dénominations soient proposées comme « langue des cités » (ou encore « français contemporain des cités » Goudaillier, 1997) et discutées (voir par exemple Trimaille et Billiez, 2007 sur la dénomination « parler jeune »), celle de « pratiques langagières » a été privilégiée car elle laisse entendre que les locuteurs sont multi-compétents et rend compte du caractère social du langage. Elle intègre les dimensions interactionnelles et identitaires[4] qui sont impliquées lors d'un échange. L'expression « jeunes de banlieue parisienne » reste insatisfaisante car homogénéisante, cette terminologie « a tendance à regrouper sous une même catégorie une réalité plurielle » (Boyer, 2012 : 43). Toutefois dans le cadre de cette étude elle réfère concrètement aux locuteurs et locutrices qui ont été enregistré·e·s et dont les productions constituent le corpus. Ils et elles sont âgé·e·s de 12 à 37 ans (tranche d'âge qui correspond à celle du MPF). Définir ce qu'est un « jeune » paraît utopique pour une notion floue et insaisissable dont les critères parfois retenus semblent aujourd'hui instables (Padis, 2005 : 17). L'âge reste une donnée biologique manipulée (Bourdieu, 1984 : 145). Quant au terme « banlieue parisienne », il désigne les villes qu'habitent et investissent par leur discours les individus qui ont contribué au corpus. La question de la dénomination

[1] https://www.ortolang.fr/market/corpora/mpf et Gadet, 2016.

[2] http://cfpp2000.univ-paris3.fr/ Pour une présentation du corpus : Branca-Rosoff *et al.* 2012 et plus récemment le n° 15 de *Corpus*.

[3] Les marques alternatives de genre, ou écriture inclusive, n'ont pas été conservées tout au long de l'article pour des raisons de lisibilité mais surtout de format. Il est évident que le chercheur peut aussi être une chercheuse, l'enquêteur une enquêtrice etc.

[4] Guerin (2017) pour compléter la notion « d'identité ».

pourrait être discutée plus longuement[5] toutefois l'objectif de cet article est centré autour de la problématique suivante : Comment récolter des données attestées permettant l'étude de ces pratiques ?

Les 5 étapes constituant la chaîne de production[6] d'un corpus peuvent être représentées sous la forme d'un schéma.

Illustration 1 : Les 5 étapes de la constitution d'un corpus de données orales

Ces étapes sont successives et bidirectionnelles puisqu'elles ne sont pas incompatibles avec une rétrospection. En effet il est possible qu'à l'écoute d'un enregistrement obtenu à la suite des étapes 1 et 2, l'enquêteur s'aperçoive que tel ou tel propos tenu semble être intéressant à développer. Il peut alors choisir de reconduire ses enquêtes et rediriger son protocole. De même lors de l'exploitation des données transcrites, il peut s'apercevoir qu'une information aurait dû être transcrite. L'objectif reste d'éviter ces retours coûteux en termes de temps et de moyens, et qui sont souvent difficilement réalisables. Il est donc important de ne pas négliger la première étape, qui vient en amont du recueil des données.

1. Étape 1 : Objectif du corpus.

L'étape qui précède le travail sur le terrain implique certaines interrogations, pour lesquelles chaque réponse doit être justifiée : quel est l'objectif de la recherche ? Quels types de données récolter ? Auprès de qui ? Comment les récolter ? Avec quel(s) outil(s) ? Comment doit se présenter l'enquêteur (paradoxe de l'observateur) ?

Ces questions se sont également posées pour la récolte des données en vue de l'analyse syntaxique sur laquelle s'appuie l'article. Pour illustration, les réponses apportées aux trois premières questions peuvent être résumées ainsi : (i) l'objectif de la recherche est de vérifier l'hypothèse, ou de l'ajuster, qu'il existerait un trait syntaxique[7] permettant de repérer une pratique langagière propre à certains jeunes de banlieue parisienne. (ii) Pour cela il faut collecter des données orales attestées auprès des locuteurs de ces façons de parler. (iii) Trois critères ont alors été retenus pour les identifier : « jeunes », « issus de milieu modeste ou populaire », « connaissant des contacts multiculturels réguliers » (Gadet, Guérin, 2016 : 286).

2. Étape 2 : Récolte des données.

Quelle méthodologie adopter pour recueillir ces productions langagières spécifiques ? Pour répondre à cette question nous nous appuierons sur la méthodologie utilisée pour le corpus personnel (proche de celle du MPF), l'objectif étant de récolter des données ordinaires et naturelles. La proximité des protagonistes a été le critère principal pour ces enregistrements. Ainsi les propos enregistrés sont en général plus spontanés, il n'y a pas ou peu d'influence provoquée par la présence d'un enquêteur

[5] Voir plus récemment Gadet 2017 : 45-48.

[6] L'expression « chaîne de production » a notamment été employée dans Traverso 2016 : 169. Les étapes présentées dans cet ouvrage recoupent celles exposées ici. Quelle que soit l'analyse finale des corpus oraux, les grandes étapes semblent équivalentes bien que des diversités existent au sein même de chaque étape (diversité au niveau des conventions de transcription par exemple).

[7] Trait syntaxique abordé à l'étape 5.

dont la position induirait une inhibition des façons de parler[8] de ces jeunes d'Île-de-France. Le lien entre l'enquêteur et l'enquêté est antérieur au recueil. Cette proximité a permis d'obtenir des enregistrements selon deux types d'interactions : les entretiens dits de « proximité » (Gadet, 2017 : 16) et les données écologiques[9] définies comme étant des « évènements discursifs non provoqués, […] enregistrés avec l'aide et la complicité d'informateurs, [dans] des activités ordinaires, avec leurs interlocuteurs familiers » (Gadet, 2017 : 17). Les enregistrements ont été effectués dans plusieurs situations. L'enquêté a été enregistré ou s'est enregistré avec des personnes avec lesquelles il entretient différents liens. Il est soit avec ses responsables de travail, soit avec sa famille, soit avec ses amis.

Le matériel utilisé est un Smartphone. Bien que la qualité obtenue avec ce type d'outil ne soit pas aussi bonne qu'avec un microphone dans un milieu insonorisé, les téléphones portables avec enregistreur vocal ont été privilégiés car ils sont plus discrets et aident ainsi aux enquêtés à faire abstraction de l'enregistrement. Ce choix a contribué à mettre à l'aise aussi bien l'enquêteur que les enquêtés puisque cet objet fait partie de l'environnement de tous.

Enfin lors de cette étape l'enquêteur doit aussi penser aux aspects juridiques et éthiques liés à la collecte de données et à l'utilisation des corpus oraux (Baude *et al*. 2006). Il doit prendre soin d'informer les personnes prenant part à l'enquête des enjeux et s'assurer de leur autorisation, sans trop influencer la qualité des productions recueillies.

3. Étape 3 : Sélection et traitement des données.

Une fois les données collectées, il faut alors les traiter. Cette troisième étape recouvre plusieurs phases.

3.1. Sélection des données

Toutes les données ne vont pas être transcrites, car l'étape de la transcription est coûteuse. Une première écoute doit donc être effectuée pour sélectionner les enregistrements « prioritaires ». Peuvent être considérés comme prioritaires les enregistrements longs et de bonne qualité sonore, c'est-à-dire avec le moins de passages inaudibles possible, pour éviter une analyse linguistique difficile voire faussée quand l'interprétation du transcripteur est souvent sollicitée. Les enregistrements où le phénomène étudié semble apparaître plus fréquemment seront transcrits en premier, puis viendra le tour de ceux qui semblent les plus intéressants pour l'ensemble de la communauté scientifique[10].

3.2. Logiciel de traitement de données[11]

Le choix du logiciel est aussi directement lié aux objectifs d'analyse et participe à la mise à disposition des données. Pour le corpus personnel, *Transcriber*[12] a été retenu car il permet un alignement son et écrit. L'interface de visualisation de la transcription (illustration 2) facilite la lecture du texte oral. La sortie des fichiers en format XML permet de convertir et d'utiliser facilement les données dans d'autres logiciels comme un concordancier du type d'*Antconc*[13]. Avant de procéder à la transcription, l'enregistrement brut peut être modifié avec un logiciel de traitement de son, tel que *Audacity*[14], pour réduire les bruits (condition écologique signifie aussi bruit environnant comme les voitures, le vent, la télévision, etc.).

---

[8] Entendre par là les façons de parler qu'ils ont habituellement entre pairs. Ces façons de parlers varient bien sûr selon le contexte interactionnel.

[9] L'entretien de proximité peut être aussi qualifié d'écologique. Les deux types sont distingués ici car dans le cas des données purement écologiques l'enquêteur est totalement absent des enregistrements. Ceci est possible lorsque l'enquêté devient alors informateur et va lui-même collecter les données.

[10] Une part de subjectivité réside malgré tout dans le choix des enregistrements à transcrire.

[11] Des articles tels que Cappeau, Gadet, Guérin, Paternostro, 2011 et plus récemment Cappeau et Gadet, 2016 traitent plus en détail du choix de ces outils et de leurs impacts sur les données et leurs analyses.

[12] http://perso.ens-lyon.fr/matthieu.quignard/Transcriber/

[13] http://www.laurenceanthony.net/software.html

[14] https://www.audacityteam.org/

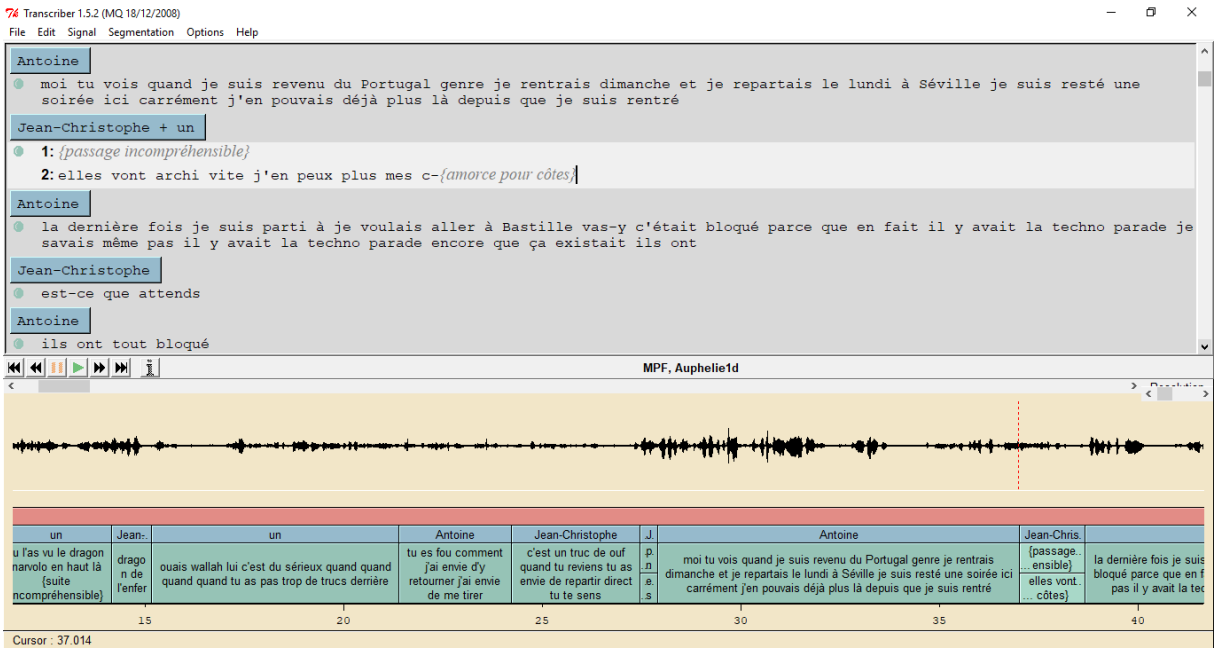

Illustration 2 : Extrait de la transcription [MPF, Auphelie1d] sur le logiciel Transcriber

3.3. Conventions de transcription

Le choix des conventions de transcription influence également la lecture des données et leurs analyses. Il répond aux objectifs de la recherche (voir par exemple les conventions adoptées pour une analyse du français parlé en interaction dans Traverso, 2016). Dans le cadre de la recherche sur laquelle s'appuie l'article, il a été privilégié l'orthographe standard pour faciliter la lecture du corpus et son exploitation. Le fait de ne pas recourir à des trucages orthographiques permet aussi d'éviter de transmettre l'interprétation ou la perception du transcripteur lors de la diffusion des données. Les conventions correspondent à celle du GARS (Blanche-Benveniste et Jeanjean, 1987 et Blanche-Benveniste, 2010). Aucun aménagement n'est fait, hormis la notation des amorces (définies comme « l'abandon d'un mot en cours de production » (Traverso, 2017 : 173) et noté par un tiret « - ») et des passages inaudibles (notés en commentaire « passage incompréhensible »). Une anonymisation du signal sonore et de la transcription est réalisée.

3.4. Les métadonnées

Les métadonnées doivent contenir le plus d'informations possible dans l'intérêt de l'analyse, dans l'optique du partage du corpus, notamment pour l'ouvrir à un plus grand champ d'analyses, et dans un souci de conservation des informations. L'organisation de ces fichiers contribue à leur pérennité.

4. Étape 4 : Rassemblement des données.

Le rassemblement des enregistrements extraits des trois sous-corpus (MPF, données personnelles et CFPP2000) invite à s'interroger sur la comparabilité des données au sein même de ces corpus et entre les corpus. En effet la proximité entre enquêteur et enquêté, dans le MPF et les données personnelles, a permis d'obtenir des enregistrements de nature différente que ceux issus d'entretiens semi-dirigés à la manière du CFPP2000. La proximité entre enquêteur et enquêté a un impact sur la langue elle-même. Le locuteur face à un inconnu modifiera sans doute sa façon de parler, afin de se rapprocher au maximum d'une norme imaginée. Il y a donc « un risque d'auto-surveillance » (Gadet et Wachs, 2015 : 39).

5. Étape 5 : Analyse.

La dernière étape est celle de l'analyse. La volonté d'approcher les « pratiques langagières des jeunes de banlieue parisienne » à travers un phénomène syntaxique a induit la constitution de ce corpus. Le phénomène syntaxique étudié est l'alternance des constructions [V. *qu-* V.] (illustré en 1) [V. Ø V.] (illustré en 2) avec les verbes dits « faibles » (Blanche-Benvensite et Willems, 2007) *croire* et *penser*.

(1) Je crois que même toi Margot tu l'aurais attrapé tu lui aurais cassé le bras [MPF, Aristide 2b]
(2) Je crois elle est au bout de sa vie [MPF, Sandrine2]

Après extraction des formes verbales *je crois* et *je pense* et leur répartition selon le type de construction (tableau 1), on observe que la construction [V. Ø V.] semble plus présente chez les jeunes de banlieue parisienne (MPF) que chez les jeunes d'un autre milieu socio-culturel comme ceux sélectionnés dans le CFPP2000 habitant Paris intramuros. Le tableau ci-dessous expose les résultats obtenus.

Verbe	Je crois		Total	Je pense		Total
Construction	V. *qu-* V.	V. Ø V.		V. *qu-* V.	V. Ø V.	
MPF	42	12	110*	143	8	246*
CFPP2000	6	0	12	55	1	76*

* L'écart entre le nombre total d'occurrences et la somme des occurrences pour chacune des constructions correspond au nombre d'occurrences des verbes en tant que recteur fort ou en construction autre.

Tableau 1 : Nombre des constructions [V. *qu-* V.] et [V. Ø V.] avec les verbes *je crois* et *je pense* dans les données issues du MPF et du CFPP2000 (au total 180 000 mots interrogés pour le MPF et 60 000 pour le CFPP2000)

Néanmoins au regard de la nature différente des données (conversations purement écologiques vs entretiens semi-dirigés, proximité vs distance, etc.) on peut se demander si le contexte interactionnel n'aurait pas une large influence pour cette variable. L'analyse des productions d'Antoine, l'enquêté principal du corpus personnel, confirme cette hypothèse. Lorsqu'il est en interaction avec ses ami·e·s et sa famille la construction [V. Ø V.] apparaît. Elle est toutefois absente des enregistrements où Antoine échange avec ses responsables de travail et lorsqu'il est en entretien formel avec l'enquêteur. De plus amples analyses sont nécessaires mais on peut d'ores et déjà poser que ce trait, qui par ailleurs peut sembler pertinent pour caractériser une des façons de parler des jeunes de banlieue parisienne, est sensible à la dimension interactionnelle.

Conclusion

Cet article avait pour objectif de rendre compte des processus, des difficultés et des devoirs (envers l'enquêté et la communauté scientifique) de l'enquêteur lors de la constitution de son corpus d'étude. Un bon nombre des facteurs qui influencent la réalisation d'un corpus a été exposé dans le cadre d'une approche des « pratiques langagières des jeunes de banlieue parisienne ». Les cinq étapes

présentées restent sensiblement identiques pour toute constitution de corpus oraux de données attestées. La constitution d'un corpus d'étude reflète toujours l'objectif et l'orientation du chercheur.

Bibliographie :

BAUDE Olivier, BLANCHE-BENVENISTE Claire, CALAS Marie-France, CAPPEAU Paul, CORDEREIX Pascal, et al. (2006), *Corpus oraux, Guide des bonnes pratiques*, Paris, CNRS.

BLANCHE-BENVENISTE Claire (2010), *Approches de la langue parlée*, Paris, Ophrys.

BLANCHE-BENVENISTE, Claire et JEANJEAN Colette (1987), *Le français parlé, transcription et édition*, Paris, Didier-érudition.

BLANCHE-BENVENISTE Claire et WILLEMS Dominique (2007), « Un nouveau regard sur les verbes « faibles » » ?, *Bulletin de la société de linguistique de Paris*, t.CII, fasc.1, p.217-254.

BRANCA-ROSOFF Sonia, FLEURY Serge, LEFEUVRE Florence (2012), *CFPP2000 Discours sur la ville, Corpus de français parlé parisien des années 2000*, http://cfpp2000.univ-paris3.fr/CFPP2000.pdf (14.04.18).

BOURDIEU Pierre (1988), « La jeunesse n'est qu'un mot », entretien avec A.-M. Métailié paru dans BOURDIEU P., *Questions de sociologie*, Edition de Minuit, p. 143-154.

BOUTET Josiane, FIALA Pierre, SIMONIN-GRUMBACH Jenny (1976), « Sociolinguistique ou sociologie du langage », *Critique*, vol. 344, p. 68-85.

BOYER Isabelle (2012), « Entre discrimination et valorisation : représentation du jeune de banlieue ou franchir ou pas les « murs » de la cité », in : *Ségrégation, normes et discrimination(s). Sociolinguistique urbaine et migrance*. (M. Lebon-Eyquem, T. Bulot, G. Ledegen, dirs.), Coll. Proximités-Sciences du langage, p. 43-59.

CAPPEAU Paul, GADET Françoise, GUERIN Emmanuelle et PATERNOSTRO Roberto (2011), « Réflexions sur les incidences de quelques aspects de la transcription outillée. », *LINX*, vol. 64-65, p. 85-100.

CAPPEAU Paul et GADET Françoise (2016), « 'Quand l'oeil écoute'... Que donnent à lire les transcriptions d'oral ? », *Actes du XXVIIe Congrès international de linguistique et de philologies romanes* (Nancy, 15-20 juillet 2013), Section 9 : Rapports entre langue écrite et langue parlée, p. 1035-1045.

TRAVERSO Véronique (2016), *Décrire le français parlé en interaction*, Paris, Ophrys.

GADET Françoise (2016), « Construire un corpus pour des façons de parler non standard : 'Multicultural Paris French'. », *Corpus*, vol. 15, p. 285-307.

GADET Françoise (coord.) (2017), *Les parlers jeunes dans l'Île-de-France multiculturelle*, Paris, Ophrys.

GADET Françoise et GUERIN Emmanuelle (2012), « Des données pour étudier la variation : petits gestes méthodologiques, gros effets. », *Cahiers de linguistique*, vol. 38/1, p. 41-65.

GADET Françoise et WACHS Sandrine (2015), « Comparer des données de corpus : évidence, illusion ou construction ? », *Langage & Société*, vol. 154, p. 33-49.

GOUDAILLIER Jean-Pierre (1997), *Comment tu tchatches ! Dictionnaire du français contemporain des cités*, Paris, Maisonneuve et Larose.

GUERIN Emmanuelle (2017), « Éléments pour une approche communicationnelle de la variation », in *La variation en question(s)*, (H. Tyne, M. Bilger, P. Cappeau, E. Guerin), Bern, Peter Lang, p. 57-73.

PADIS, Marc-Olivier, 2005, « L'émergence de la catégorie jeunesse », *Les Cahiers de Profession Banlieue*, mars 2005, p.12-27.

TRIMAILLE Cyril et BILLIEZ Jacqueline (2007), « Pratiques langagières de jeunes urbains : peut-on parler de "parler" ? », in : *Les français en émergence*, (E. Galazzi, C. Molinari éds), Bern, Peter Lang, p. 95-109.