CORPUS

Corpus

22 | 2021
Du recueil à l'outillage des corpus oraux : comment accéder à la variation ?

Le corpus comme portail pour l'étude de la variation (socio)linguistique

Shana Poplack

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/5422
DOI : 10.4000/corpus.5422
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Shana Poplack, « Le corpus comme portail pour l'étude de la variation (socio)linguistique », *Corpus* [En ligne], 22 | 2021, mis en ligne le 10 février 2021, consulté le 16 février 2021. URL : http://journals.openedition.org/corpus/5422 ; DOI : https://doi.org/10.4000/corpus.5422

Ce document a été généré automatiquement le 16 février 2021.

Le corpus comme portail pour l'étude de la variation (socio)linguistique*

Shana Poplack

1. Introduction

1 Cet article détaille les principes et pratiques de gestion de données en vigueur au Laboratoire de sociolinguistique de l'Université d'Ottawa (le Labo ; http://www.sociolinguistique.uottawa.ca/laboratoire.html), centre où sont stockées des centaines d'heures d'enregistrements de parler spontané. Le Labo, que je dirige depuis sa fondation en 1982, a pour mandat de promouvoir l'étude de la langue courante et, en particulier, de la variation et du changement linguistiques, notamment dans les contextes minoritaires et bilingues. La réalisation de ce mandat implique l'étude systématique de corpus oraux construits selon des principes scientifiques, en se fondant sur la théorie et les méthodes de la sociolinguistique variationniste. Un grand souci de scientificité, en particulier la capacité de répliquer et de valider nos résultats empiriques, sous-tend tous nos travaux.

2 Le Labo abrite 19 grands corpus de parler spontané en diverses langues, dont huit compilations de discours bilingue comportant des emprunts et des alternances de codes entre différents couples de langues[1], tous construits par notre équipe. Parmi ces corpus figurent de volumineuses banques de données du français parlé au Canada sur une période d'un siècle et demi (323 locuteurs ; plus de 4 000 000 mots), de l'anglais parlé dans des communautés de la diaspora afro-américaine (124 locuteurs ; 223 heures) et de l'anglais parlé au Québec avant et après l'adoption de la Charte de la langue française qui en a fait une langue minoritaire (164 locuteurs ; 2 500 000 mots). On y trouve aussi les *Archives de sociolinguistique*, composées de plus de 700 entrevues recueillies dans la région de la capitale fédérale entre 1982 et 2018 par des générations d'étudiants dans le cadre du cours de Dialectologie urbaine. Le Labo renferme également trois importants

corpus écrits reflétant l'oral (*Ottawa Repository of Early African American Correspondence* (537 lettres personnelles rédigées entre 1790 et 1865 par des Afro-Américains semi-lettrés ; Van Herk et Poplack, 2003) ; *Ottawa Grammar Resource on Early Variability in English* (98 grammaires publiées entre 1577 et 1930 ; Poplack *et al.*, 2002) ; *Recueil historique des grammaires du français* (163 grammaires publiées entre 1530 et 1999 ; Poplack *et al.*, 2015)).

3 Une grande partie de ce travail a été amorcée au début des années 1980, bien avant que la construction de corpus et la gestion de données ne deviennent des sujets d'actualité en dehors des cercles de la sociolinguistique variationniste. De ce fait, certaines des méthodes décrites ici paraîtront plutôt archaïques comparées aux normes actuelles. Néanmoins, 40 ans plus tard, ces ressources ont conservé leur intérêt et leur utilité, comme en témoignent les nombreux articles, livres, thèses, dissertations, conférences, ateliers et autres travaux produits par notre équipe et par nos collaborateurs et collègues[2]. Cela s'explique par le fait que tous ces corpus ont été bien préservés et restent exploitables, consultables, et dans la mesure où les contraintes déontologiques le permettent, partageables. Les sections suivantes décrivent comment nous avons abordé les tensions omniprésentes entre l'idéal et le réalisable pour parvenir à un tel résultat.

2. Constitution de corpus

2.1. Le corpus au profit d'une question de recherche

4 Conscients que le matériel linguistique à la disposition du linguiste décide en grande partie de ce qui peut faire l'objet d'étude, nous amorçons notre démarche par le recueil de données, ce qui soulève l'inévitable question : quoi recueillir et auprès de qui ? Les corpus du Labo sont d'abord et avant tout conçus comme des archives de réponses potentielles à des problèmes de recherche précis. Fidèles au mandat de la sociolinguistique d'étudier les enjeux linguistiques qui revêtent une importance particulière pour la société, nous privilégions ceux qui émergent du discours public, surtout lorsqu'ils présentent un intérêt linguistique théorique.

5 Un exemple notable est le *Corpus du français parlé à Ottawa-Hull* (OH ; Poplack, 1989), pierre angulaire d'un projet toujours en cours qui vise à déterminer dans quelle mesure le statut minoritaire a une incidence sur la perméabilité à l'influence d'une langue majoritaire. La nature, l'étendue et même l'existence des changements provoqués par le contact des langues ont longtemps suscité la controverse en linguistique (p. ex. Heine et Kuteva, 2005 ; Poplack et Levey, 2011 ; Thomason, 2001 ; Winford, 2003). C'est aussi une préoccupation bien ancrée chez les francophones du Canada, qui redoutent que l'omniprésence de l'anglais détruise l'intégrité structurelle du français. De telles craintes, jusqu'à tout récemment peu validées de façon empirique, ont dicté notre méthodologie : le corpus OH (120 locuteurs ; 3 500 000 mots) est stratifié en fonction du statut majoritaire ou minoritaire du français ainsi que de l'intensité du contact au niveau local. À l'intérieur de ce cadre, le choix des locuteurs s'est fait de façon aléatoire, ajoutant une rare représentativité à l'échantillon. Un autre corpus (*Le français en contexte : milieux scolaire et social* ; Poplack, 2015) a été conçu à l'origine pour mettre à l'épreuve la perception populaire que les jeunes francophones ne parlent pas un « bon » français parce que leurs enseignants ne le maîtrisent pas eux-mêmes. Il met en

jeu 166 lycéens et 24 enseignants de français. Nous avons recueilli plus d'un million de mots dans la même région ciblée 25 ans plus tôt par le corpus OH, conférant ainsi une dimension de temps réel à l'analyse. Les *Récits du français québécois d'autrefois* (Poplack et St-Amand, 2009) est un corpus constitué d'enregistrements sonores réalisés auprès de Québécois de milieux ruraux par des folkloristes dans les années 1940 et 1950. Nous avons ciblé 37 locuteurs nés entre 1846 et 1895 (524 000 mots). En conjonction avec le corpus OH, il nous permet d'étendre la portée temporelle de l'étude du changement à un siècle et demi en temps apparent (et à plus de 60 ans en temps réel), une durée virtuellement inégalée pour l'étude de l'oral. Les *Récits* constituent également un repère temporel antérieur au contact intense avec l'anglais, élément crucial à toute étude du changement dû au contact (Poplack et Levey, 2011).

6 Cette approche comparative et diachronique sous-tend également la constitution des corpus de l'anglais vernaculaire afro-américain (AVAA) des XVIII[e] et XIX[e] siècles[3], nés du débat sur le précurseur de l'AVAA (créole ou dialecte d'anglais). Ces données permettent de répondre aux préoccupations des locuteurs natifs concernant la « qualité » de leur langue, tout en respectant l'impératif de se reporter à un stade antérieur pour étudier les origines. Les corpus d'AVAA misent sur les enregistrements sonores synchroniques du parler de descendants des premiers colons de trois isolats de la diaspora afro-américaine établis entre 1783 et 1824. En raison de leur isolement depuis, ces parlers reflètent un stade antérieur de la langue, fournissant ainsi la chance inouïe de reconstruire l'ancêtre de l'AVAA contemporain. Nous y parvenons en comparant le fonctionnement de certaines structures grammaticales entre isolats et par rapport à des variétés-repères pertinentes (Poplack, 2000 ; Poplack et Tagliamonte, 2001).

7 Pour une liste complète des collections du Labo, consulter http://www.sociolinguistique.uottawa.ca/recherche.html et les références afférentes. Il s'agit pour la plupart de corpus « non conventionnels » (Beal *et al.*, 2007 ; Poplack, 2007), dans la mesure où ils ont été conçus en vue d'aborder un enjeu particulier, en plus de documenter le parler d'une communauté définie en termes extralinguistiques, comme il est plus souvent le cas en (socio)linguistique. Grâce aux méthodes de collecte de données détaillées ci-dessous, ces banques de données se prêtent aisément à l'étude, tant synchronique que diachronique, de n'importe quel élément linguistique, du moment qu'il se produise à l'oral et puisse être saisi à partir d'un enregistrement sonore. De plus, en vertu des critères qui sous-tendent leur constitution, ces corpus offrent la rare possibilité d'interpréter le comportement de chaque trait linguistique étudié en fonction d'une question de recherche plus large, un avantage inestimable des corpus construits selon de tels principes.

2.2. Les données ciblées

8 En tant que sociolinguistes, notre préoccupation principale est l'étude de la parole spontanée, et en particulier de la variabilité qui la caractérise. Cette variabilité implique généralement une alternance entre des formes ratifiées et leurs contreparties non standard, qui sont souvent stigmatisées, sinon franchement condamnées. Ces dernières sont emblématiques du *vernaculaire*, considéré comme la forme la plus régulière et systématique de la langue (Labov, 1966/2006). Il n'est pas si facile de recueillir des témoins du vernaculaire au cours du processus formel de collecte de données. Il faut d'abord et avant tout créer une situation dans laquelle son utilisation

est jugée appropriée par le locuteur. La tâche est relativement simple quand il s'agit de l'étude de réseaux sociaux modestes, qui impliquent souvent l'observation à long terme et la familiarisation connexe avec leurs membres ; elle se révèle nettement plus complexe lorsque l'analyste cherche également à constituer un échantillon statistiquement représentatif et quantitativement suffisant. Nos corpus principaux sont composés de plus de 100 participants, pour la plupart sans lien les uns avec les autres[4]. Ce que de telles études à grande échelle gagnent en extension, elles le perdent souvent en profondeur. Le parler qui en résulte ne s'éloigne que rarement des pôles plus formels du continuum stylistique, où les traits linguistiques d'intérêt sont rares ou simplement absents ; c'est là un obstacle de taille à l'étude du vernaculaire.

9 En réponse à ce problème, nous avons adopté les méthodes d'inspiration ethnographique développées par Labov et ses collaborateurs pour résoudre le *Paradoxe de l'observateur* (Labov, 1972). Celles-ci invitent à une conversation qui se rapproche davantage du parler de tous les jours que du registre généralement réservé aux entretiens en face-à-face. L'outil méthodologique privilégié est l'*entrevue sociolinguistique* (Labov, 1984). Antithèse du protocole d'entrevue conventionnel, il s'agit d'un guide pour encourager le participant à s'exprimer de façon informelle en proposant une vaste gamme de sujets de conversation eux-mêmes de nature informelle. Pour réduire l'effet du contexte de l'entrevue, notamment du rapport de pouvoir entre l'intervieweur et le locuteur, on encourage ce dernier à contrôler l'inclusion et l'exclusion des sujets de conversation, en minimisant les interventions de la part de l'intervieweur. Les seules exceptions concernent la collecte de métadonnées (section 3.1), qui se fait vers la fin de chaque séance d'enregistrement.

10 Ces efforts, décrits en détail ailleurs (p. ex. Poplack, 1989 ; Poplack *et al.*, 2006), ont donné lieu à une mine d'enregistrements du discours spontané, dont la durée varie de une à cinq heures par participant et qui comprennent des récits d'expériences personnelles, des discussions en groupe et d'autres modes de conversation hautement informels. On y trouve bon nombre des variantes vernaculaires si convoitées, en plus du discours soigné, plus facilement accessible. Dans la mesure du possible, nous comptons sur des membres de la communauté ciblée pour s'acquitter du travail sur le terrain. Ceux-ci reçoivent à cette fin une formation en vue d'administrer l'entrevue sociolinguistique, qui se trouve grandement enrichie par leur expertise personnelle sur les mœurs de la communauté.

11 L'*African Nova Scotian English Corpus* (Poplack et Tagliamonte, 1991) témoigne sans doute de notre plus grande réussite dans ce genre d'entreprise. Il s'agit d'une vaste compilation de discours recueillis par des membres de collectivités très soudées qui sont diglossiques en anglais canadien standard et en AVAA. Si ce n'était du statut local – et de la grande compétence – de ces travailleurs de terrain, la plupart des traits grammaticaux vernaculaires que nous avons étudiés en détail depuis (par exemple, l'élision de la copule (Walker, 2000), le marqueur de négation *ain't* (Howe et Walker, 2000) ou le manque d'accord verbal (McOrmond-Arenja, 2020 ; Poplack et Tagliamonte, 1989)), ne seraient tout simplement pas apparus pendant les séances d'enregistrement. En effet, là où coexistent l'insécurité linguistique et un certain degré de diglossie, les formes « mal vues » sont généralement évitées avec les interlocuteurs externes à la communauté. D'où l'impératif de créer des conditions appropriées pour la collecte de données. Les corpus du Labo recensent d'autres exemples de ces phénomènes langagiers convoités mais fugaces, notamment l'alternance de code multimot (Poplack,

1985) et le registre vernaculaire du français canadien parfois (péjorativement) appelé joual.

3. Traitement des données

3.1. L'importance des métadonnées

12 Comme le locuteur est à la fois source principale de variabilité inhérente à la langue et agent clé du changement, nous tentons, dans la mesure du possible, d'intégrer ses caractéristiques pertinentes à nos analyses linguistiques. Cette section décrit les métadonnées développées au Labo afin de faciliter cette initiative. Tout d'abord, après chaque séance d'enregistrement, les intervieweurs remplissent un rapport d'entrevue dans lequel ils fournissent des renseignements démographiques, sociologiques et linguistiques au sujet du participant et de son entourage immédiat. À ce moment-là, l'interaction est anonymisée par l'attribution d'un pseudonyme et d'un numéro d'identification temporaires qui sont soigneusement associés à chaque enregistrement, à chaque formulaire de rapport d'entrevue et à tout autre document y afférant[5]. Ces identifiants sont ensuite transcrits sur la liste maîtresse de métadonnées rattachée au corpus et mise à jour au fil de l'acquisition des données. Une fois la construction du corpus terminée, chaque participant se voit attribuer un pseudonyme et un numéro d'identification permanents, qui le rattachent systématiquement à l'ensemble de ses données.

13 On obtient ainsi un inventaire détaillé de caractéristiques potentiellement pertinentes. L'importance relative de ces caractéristiques fluctue d'une variable à l'autre et d'une communauté à l'autre : une variable peut être surtout conditionnée socialement tandis qu'une autre sert de marqueur stylistique et qu'une troisième affiche un profil qui varie selon la communauté. Les caractéristiques sociodémographiques classiques (l'âge, le sexe, le quartier de résidence, le niveau d'éducation ou le statut socioéconomique approximatif) sont toujours prises en considération, mais nous tenons également compte des enjeux locaux là où il est possible de cerner leur rôle de façon objective. Ainsi, l'attitude des locuteurs à l'égard d'une langue majoritaire peut s'avérer explicative dans l'analyse de la variabilité dans un contexte minoritaire (Poplack *et al.*, 2006) ; la maîtrise rapportée de la L2 ou la propension à alterner les codes ou à emprunter des mots pourraient être significatives dans les communautés bilingues (Poplack, 1989 ; 2018). Ces facteurs sont opérationnalisés par l'entremise d'un codage qui s'intègre facilement aux analyses statistiques afin d'évaluer leur contribution relative aux choix de variantes que font les locuteurs.

14 Un dernier type de métadonnées est fourni par l'« article de corpus » décrivant la raison d'être du projet global, ainsi que des détails méthodologiques pertinents concernant le type de communauté, les modalités de collecte de données, les critères de sélection de l'échantillon, la description des locuteurs et des données, les protocoles de transcription, etc. (Poplack, 1989 ; Poplack et St-Amand, 2009 ; Poplack *et al.*, 2006). Puisque les données comme les résultats ne peuvent être pleinement interprétés que dans le contexte des normes de la communauté au sein de laquelle ils ont été recueillis, les usagers éventuels sont tenus de confirmer qu'ils ont lu ces publications avant de recevoir l'autorisation d'accéder aux corpus.

3.2. Représentation fidèle du parler variable

15 Avant de décrire les protocoles de transcription appliqués aux corpus du Labo, il faut rappeler qu'une grande partie des données a été recueillie des décennies avant l'avènement des outils d'annotation disponibles aujourd'hui (p. ex. Beal *et al.*, 2007 ; Baude et Dugua, 2016 ; Eshkol-Taravella *et al.*, 2012). En comparaison, la solution que nous avons adoptée – la transcription manuelle en orthographe standard – peut paraître plutôt simpliste. Pourtant, même si des méthodes plus sophistiquées ont gagné en popularité au fil des ans, nous nous en sommes largement tenus à nos protocoles originaux, même pour les corpus subséquents, ayant constaté par expérience qu'ils répondent bien à nos besoins. Notre conception de l'annotation comme portail vers l'analyse, plutôt que comme fin en soi, a dicté ce choix.

16 Tout d'abord, une fois que nous avons personnellement constaté le formidable investissement de temps et de fonds requis pour monter *ex nihilo* un grand corpus basé sur des principes raisonnés, nous avons décidé de répartir nos ressources limitées de façon à faire prévaloir l'analyse linguistique des données plutôt que leur gestion, sans toutefois sacrifier celle-ci à l'excès. Cet état de fait a motivé notre décision d'expédier la transcription et de multiplier les étapes de correction (entre trois et six, manuelles et automatisées, selon les corpus).

17 Parvenir à une représentation à la fois fidèle et cohérente s'avère particulièrement ardu dans le cas du parler spontané, qui est caractérisé par la variabilité inhérente, impliquant souvent de nombreuses formes non standard – parfois, comme dans notre cas, dans plus d'une langue. Mais la valeur d'un système d'annotation est proportionnelle à sa capacité de servir les objectifs de l'étude. Nous avons mentionné plus haut que les corpus du Labo permettent d'étudier une grande variété de phénomènes linguistiques. Certains de ces phénomènes sont connus au départ, mais la plupart émergent au fur et à mesure que le projet évolue. Devant l'impossibilité de prévoir leur émergence, nous avons conclu qu'il serait déraisonnable, voire impossible, de tenter de les représenter tous à l'étape de la transcription. À titre d'exemple, comme nos recherches portent principalement sur l'analyse de la variabilité morphosyntaxique, nous avons choisi d'ignorer la multitude de variantes phonétiques présentes dans les enregistrements et de ne conserver que la variation morphosyntaxique pertinente, évitant ainsi de multiplier inutilement les entrées et d'entraver le repérage.

18 Le choix du protocole de transcription s'est fait en fonction de notre objectif premier : construire une concordance informatisée qui permet un rappel maximal des données, ce qui suppose un haut degré de cohérence de la transcription. Voilà pourquoi nous avons adopté une solution orthographique, décrite en détail dans Poplack (1989). Notre stratégie générale consiste à rendre les variantes résultant de processus phonétiques ou phonologiques en orthographe standard, peu importe leur réalisation réelle (p. ex. <ing> tant pour la variante vélaire [ɪŋ] que pour l'alvéolaire [ɪn] (dans l'exemple en (1)), mais à rendre les variantes morphophonologiques et morphosyntaxiques telles qu'elles ont été produites (p. ex. <trunk> ou <trunks> en (2) selon si le morphème pluriel [s] est élidé ou non).

(1) And I said, "If things don't change around here, I'm getting out of here." (QEC. 037.630)[6]

(2) That man had two trunks. Two trunk full of gold and silver and everything. Two trunk, big trunks. Full of gold and silver. (ANSE.NP.030.1323)

19 Nos modalités de transcription correspondent *grosso modo* aux conventions orthographiques de la langue correspondante, sauf là où elles contreviennent à nos critères d'accessibilité.

20 La transcription exige un effort décisionnel continu, surtout lorsque les protocoles adoptés combinent l'annotation et un minimum d'analyse, comme dans notre cas. Ce processus a été grandement facilité par le fait que nos équipes de transcription étaient composées de linguistes de formation.

3.3. Correction

21 La transcription a pour but de refléter fidèlement ce qui a été dit, y compris toute manifestation de la variabilité morphosyntaxique y afférant. Comme le confirmeront ceux qui ont déjà travaillé avec des données de parler spontané, c'est sans doute l'étape la plus laborieuse de la construction de corpus. En raison de notre stratégie de saisir les données rapidement et de nos exigences de repérabilité, l'élaboration d'un système de correction efficace s'imposait. Le nôtre comprenait plusieurs étapes dont des tours manuels (à partir de la réécoute des enregistrements audio) et des tours semi-automatisés (basées sur listes de mots et de concordances). L'échange de documents entre correcteurs a renforcé la fiabilité des transcriptions. Un suivi sur tableur a permis d'assurer l'exécution de toutes les phases de correction sans duplication d'efforts. Il en résulte un ensemble de corpus à peu près exempts d'erreurs qui peuvent être utilisés en toute confiance pour étudier maintes phénomènes morphosyntaxiques et lexicaux sans recours aux enregistrements audio d'origine. De plus, un protocole de transcription d'une telle simplicité peut aisément s'adapter à d'autres outils (p. ex. les concordanciers (section 4.1) ou les logiciels d'alignement forcé (Mielke, 2013)).

4. Analyse de la variation linguistique

4.1. Repérage

22 Les généralisations que font les variationnistes concernant le langage découlent typiquement d'analyses quantitatives à grande échelle du comportement linguistique réel. Selon la taille du corpus et la fréquence du phénomène linguistique ciblé, nos recherches peuvent porter sur quelques centaines d'occurrences (p. ex. les propositions relatives en anglais (N = 814) ; Lealess et Smith, 2011) ou sur des dizaines de milliers d'entre elles (p. ex. l'expression de la négation en français (N = 85 447) ; Poplack, 2015). Il est donc essentiel d'automatiser le traitement des données. Comme nous l'avons expliqué à la section 3.2, l'étiquetage de nos corpus se limite à identifier la langue et le locuteur. Pour repérer les données d'intérêt, nous comptons sur les concordanciers. Peu de concordanciers disponibles répondent à l'ensemble de nos besoins, notamment à la nécessité d'associer chaque mot au locuteur qui l'a produit, tout en excluant de l'analyse les données d'individus ne faisant pas partie de l'échantillon. Il faut aussi éviter que les éléments accessoires (p. ex. les métadonnées, les indications extralinguistiques comme « (rires) ») soient comptés dans les calculs appliqués aux phénomènes linguistiques. Notre outil de prédilection est Concorder X (Edwards, 2006), un concordancier développé au Labo pour répondre à nos exigences. Ici aussi, le format simple du corpus annoté permet d'adapter les données selon les exigences du logiciel

sans grandes modifications. Concorder X est un outil polyvalent qui crée efficacement des listes de mots et des concordances selon différents paramètres (p. ex. par ordre alphabétique ou selon la fréquence) tant pour un seul locuteur que pour le corpus entier ou un sous-ensemble de celui-ci. Ces fonctionnalités réduisent considérablement le temps requis pour repérer et extraire les données recherchées. La concordance affiche chaque élément lexical sous forme de mot-clé entre les contextes linguistiques le précédant et le suivant, en plus d'identifier le locuteur et l'adresse du mot dans la transcription. En cliquant sur le mot-clé, l'utilisateur accède à l'emplacement du mot dans le corpus et à son contexte d'origine en entier.

23 Les analyses variationnistes ont souvent pour but de déterminer pourquoi une variante d'une variable est choisie plutôt qu'une autre dans un *contexte variable* (point où les variantes alternent sans changer de valeur référentielle) préalablement défini. L'entrée pour chaque occurrence dans la concordance contient généralement suffisamment d'informations pour permettre à l'analyste d'en capter les facteurs potentiellement explicatifs (p. ex. la polarité de l'énoncé, la personne grammaticale, le positionnement dans la phrase, etc.). À noter cependant que l'extraction à partir d'un repère lexical risque de relever un surplus d'occurrences qui débordent du contexte variable. Ainsi, en cherchant « que » pour localiser les contextes du subjonctif, on finira avec l'ensemble des propositions subordonnées ; la recherche de « si » fera apparaître non seulement les protases hypothétiques, mais aussi les propositions comparatives. Les cas non pertinents doivent être identifiés et éliminés manuellement. Le repérage des occurrences est également compliqué par le fait que de nombreux mots grammaticaux (p. ex. « que ») sont souvent carrément supprimés à l'oral, tout comme le sont les sujets, les copules et les prépositions, pour ne nommer que ceux-là. Certaines de ces formes élidées constituent des variantes de la variable à l'étude, et doivent donc être considérées parallèlement à leurs homologues explicites. Le repérage doit donc s'effectuer en combinant la recherche automatisée (pour les formes ayant des représentations lexicales) et l'extraction manuelle (pour les éléments nuls et les variables syntaxiques comme les stratégies de formation de propositions relatives et la variation dans l'ordre des mots). L'extraction manuelle est sans contredit extrêmement exigeante, surtout dans le cas de grands corpus, mais elle présente l'avantage de permettre aux chercheurs de relever l'ensemble des variantes d'une variable donnée, condition *sine qua non* de l'analyse variationniste. Cet ensemble peut comprendre des variantes qui n'ont pas été reconnues ou identifiées au départ, comme le choix du conditionnel ou de l'imparfait dans les contextes qui demandent théoriquement le subjonctif, ou l'absorption de la préposition dans les propositions relatives françaises. Le repérage manuel oblige aussi l'analyste à se (re)familiariser continuellement avec les données analysées, données que le degré de détail de l'annotation rend proportionnellement beaucoup plus abstraites. Ce faisant, nous souscrivons à un autre principe fondamental du paradigme variationniste, à savoir que la variation linguistique doit être étudiée dans le contexte où elle se produit.

4.2. Codification

24 Quelle que soit la méthode utilisée pour les repérer, les occurrences extraites sont ensuite codées en fonction d'une série de facteurs (eux-mêmes des matérialisations d'hypothèses sur ce qui motive le choix des variantes) en vue de l'analyse statistique. Le codage des données commence par la transcription des occurrences pertinentes

directement dans des tableurs Excel. Excel offre de nombreuses fonctionnalités (filtrage, tri, tabulation, masquage des colonnes, comptage, etc.) qui facilitent le codage et améliorent sa fiabilité. Les séquences de codes résultantes sont alors soumises à l'analyse statistique afin de déterminer leur signification, leur importance relative et la direction de leurs effets. Les résultats constituent la base de nos analyses.

4.3. Au-delà du portail

25 L'utilité d'un corpus se mesure en grande partie par la polyvalence de ses applications. Les corpus du Labo relèvent le défi ; ils se prêtent à l'étude d'une grande variété de questions théoriques, dont beaucoup ont déjà fait l'objet de nos recherches, par exemple, le comportement des différentes manifestations du contact linguistique (emprunt lexical, alternance de codes, convergence grammaticale) (Poplack, 2008 ; 2018 ; Poplack et Levey, 2011), les modalités du changement linguistique (au long de la vie (Poplack et Lealess, 2009), provoqué par le contact (Leroux et Jarmasz, 2006 ; Poplack *et al.*, 2012)), la résistance des isolats linguistiques (Adams, 2005 ; Petrik, 2005 ; Poplack et Tagliamonte, 2010 ; Yoshizumi, 2006), le rôle des médias (Poplack et Dion, 2007), la grammaticalisation (en anglais (Poplack et Tagliamonte, 1996 ; 2000), en français (Poplack, 2011) et dans les langues romanes (Poplack *et al.*, 2018)), le maintien des langues ancestrales (Budzhak-Jones et Poplack, 1997), la tension entre la langue prescrite et la langue parlée (Poplack, 2015 ; Poplack *et al.*, 2015 ; Poplack *et al.*, 2002) et les origines de l'AVAA (Poplack, 2000 ; Poplack et Tagliamonte, 2001), pour ne nommer que celles-ci. Les variables linguistiques exploitées pour éclairer ces questions comprennent des phénomènes aussi disparates que l'élision de la copule (Walker, 2000), l'alternance des cas (Sankoff *et al.*, 1990), les structures interrogatives (Elsig, 2009 ; Van Herk, 2000), la variation dans l'ordre des mots (Toth, 2014), l'échouage de la préposition (Poplack *et al.*, 2019), les stratégies de formation des propositions relatives (Lealess et Smith, 2011 ; Tottie et Harvie, 2000), l'alternance des auxiliaires (Willis, 2000), l'expression variable de la référence au présent (Walker, 2001), au passé (Leroux, 2005 ; Tagliamonte, 1991 ; Van Herk, 2002) et au futur (Poplack et Dion, 2009 ; Poplack et Tagliamonte, 2000 ; Torres Cacoullos et Walker, 2009), la variation modale (Poplack, 2001 ; Poplack *et al.*, 2013 ; St-Amand, 2002), l'assignation du genre (Klapka, 2002), le marquage du pluriel (Tagliamonte *et al.*, 1997) et bien d'autres encore.

5. Consultation et préservation des données

5.2. Considérations d'ordre déontologiques

26 Toutes les données archivées au Labo ont été recueillies, traitées et entreposées conformément aux attentes déontologiques des organismes subventionnaires concernés et du Comité d'éthique de la recherche de l'Université. La seule dérogation concerne l'obtention du consentement éclairé avant d'entreprendre la collecte de données. Comme on pourrait s'y attendre, amorcer une interaction en présentant les détails linguistiques du projet et en demandant au participant de lire, discuter et signer les formulaires de consentement va à l'encontre de la création d'une atmosphère favorable au parler informel, et encore moins au vernaculaire. Nous expliquons plutôt le but de l'entrevue d'abord en termes généraux, sans manquer de signaler notre intérêt pour la langue, puis nous obtenons le consentement éclairé en faisant remplir le

formulaire de décharge immédiatement après la séance d'enregistrement. Nous l'avons déjà mentionné, plusieurs mécanismes assurent la confidentialité des données. L'identité des participants est anonymisée au moyen de pseudonymes et de numéros de locuteur ; les données fournies, tant enregistrées que transcrites, sont conservées dans des lieux sécurisés sous la surveillance de la coordonnatrice de recherche du Labo. En raison de la nature personnelle d'une grande partie des données ainsi que des différentes exigences déontologiques auxquelles elles sont assujetties, la consultation des données brutes se fait sur place, sous certaines conditions que l'utilisateur s'engage à respecter, notamment :

a. Aucune information permettant d'identifier les locuteurs ne pourra figurer dans un article, publié ou non, ou dans une communication qui utilise les données du corpus.
b. Le contenu des corpus ne servira pas à poser un jugement sur les opinions, la personnalité ou la langue du locuteur.
c. Les propos tirés des corpus seront cités verbatim et uniquement dans le but d'illustrer un point linguistique, et le contenu de toute citation devra satisfaire aux conditions (a) et (b) ci-dessus.

27 De telles précautions sont tout particulièrement importantes lorsque la variété linguistique en question est non standard ou socialement stigmatisée, comme c'est le cas pour une grande partie des données conservées au Labo.

5.2. Identification de la provenance des énoncés cités

28 L'identification de la provenance des données linguistiques n'est pas encore pratique courante. L'utilisateur des corpus du Labo doit s'engager à citer non seulement le corpus d'où les données sont extraites, mais aussi le locuteur qui les a produites. Tout énoncé reproduit dans une publication ou une présentation doit être attribué à sa source en spécifiant le nom du corpus, le numéro du locuteur et l'adresse de l'extrait[7]. Ces exigences rendent hommage à la contribution indispensable des participants, tout en facilitant la vérification des données et des affirmations connexes. Ceci augmente ainsi la reproductibilité et l'intégrité de toute étude qui s'appuie sur les données visées.

Épilogue

29 Dans le climat disciplinaire actuel, la recherche empirique que permettent les corpus est souvent dénigrée ou considérée comme théoriquement peu intéressante. En dehors du domaine de la sociolinguistique variationniste, les chercheurs sont rarement (sinon jamais) crédités pour les efforts titanesques déployés pour recueillir, transcrire, organiser et partager les vastes quantités de données de parole spontanée qui constituent bon nombre de corpus. Au contraire, ils sont souvent fustigés pour les distributions bizarres, les cases vides et les quantités parfois sous-optimales de variantes rares qui caractérisent la parole spontanée. Il arrive souvent que des revues de linguistique de pointe, considérées comme porte-parole du domaine, rejettent ou demandent une révision en profondeur des travaux quantitatifs rapportant des distributions éparses ou disproportionnées, même lorsque l'analyste a systématiquement parcouru de vastes corpus pour en extraire toutes les occurrences pertinentes. Les linguistes habitués à une analyse minutieuse de la langue parlée entendent que ces répartitions inégales des données constituent la règle plutôt que

l'exception. La méconnaissance généralisée du parler et le penchant croissant à le remplacer par des analogues plus accessibles (p. ex. le « langage » internet), dont on ne connaît pas vraiment la provenance, ont contribué à masquer ces caractéristiques fondamentales. Il est à espérer que les pratiques décrites dans cet article, pour la plupart au fondement même de la méthodologie de la sociolinguistique variationniste depuis sa création il y a plus d'un demi-siècle, contribuent à contrebalancer ce déséquilibre.

BIBLIOGRAPHIE

Adams J. (2005). *Concord Variation, Convergence, and Quebec English : 'There's Lots of Things to Consider'*. Université d'Ottawa. Mémoire de maîtrise.

Bailey G., Maynor N. & Cukor-Avila P. (1991). *The Emergence of Black English : Texts and Commentary*. Amsterdam/Philadelphia : John Benjamins.

Barysevich A. (2012). *Variation et changement lexicaux en situation de contact de langues*. University of Western Ontario. Thèse doctorale.

Baude O. & Dugua C. (2016). « Les ESLO, du portrait sonore au paysage digital », *Corpus* 15 : 29-56.

Beal J., Corrigan K. & Moisl H. (2007). *Creating and Digitizing Language Corpora : Synchronic Databases*. Houndmills : Palgrave-Macmillan UK.

Budzhak-Jones S. & Poplack S. (1997). « Two generations, two strategies : The fate of bare English-origin nouns in Ukrainian », *Journal of Sociolinguistics* 1(2) : 225-258.

Edwards J. (2006). *Concorder X : Program and Documentation*. Ottawa : Laboratoire de sociolinguistique de l'Université d'Ottawa.

Elsig M. (2009). *Grammatical Variation Across Space and Time : The French Interrogative System*. Amsterdam/Philadelphia : John Benjamins Publishing.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C. & Tellier I. (2012). « Un grand corpus oral 'disponible' : le corpus d'Orléans 1968-2012 », *Ressources linguistiques libres*, *TAL* 52(3) : 17-46.

Heine B. & Kuteva T. (2005). *Language Contact and Grammatical Change*. Cambridge University Press.

Howe D. & Walker J.A. (2000). « Negation and the creole-origins hypothesis : Evidence from early African American English », dans Poplack S. (éd.), *The English History of African American English*. Oxford & Malden : Blackwell Publishers, 109-140.

Kastronic L. (2016). *A Comparative Variationist Approach to Morphosyntactic Variation in Contemporary Hexagonal and Quebec French*. Université d'Ottawa. Thèse doctorale.

Klapka L. (2002). *Étude comparative : l'accord du genre en français québécois au XIXe et au XXe siècles*. Université d'Ottawa. Mémoire de maîtrise.

Labov W. (1966/2006). *The Social Stratification of English in New York City*. 2e édition. Cambridge : Cambridge University Press.

Labov W. (1972). *Sociolinguistic Patterns*. Philadelphia : University of Pennsylvania Press.

Labov W. (1984). « Field methods of the project on linguistic change and variation », dans Baugh J. & Sherzer J. (éd.), *Language in Use*. Englewood Cliffs : Prentice Hall, 28-54.

Lealess A.V. (2014). *« J'ai tout le temps eu de misère » : A Variationist Study of Adverb Placement in Quebec French*. Université d'Ottawa. Thèse doctorale.

Lealess A.V. & Smith C. (2011). « Assessing contact-induced language change : The use of subject relative markers in Quebec English », *Cahiers linguistiques d'Ottawa* 36 : 20-38.

Leroux M. (2005). « Past but not gone : The past temporal reference system in Quebec French », *Penn Working Papers in Linguistics (Selected Papers from NWAV 33)* 11(2) : 119-131.

Leroux M. & Jarmasz L.G. (2006). « A study about nothing : Null subjects as a diagnostic of convergence between English and French », *Penn Working Papers in Linguistics (Selected Papers from NWAV 34)* 12(2) : 1-14.

Levey S., Groulx K. & Roy J. (2013). « A variationist perspective on discourse-pragmatic change in a contact setting », *Language Variation and Change* 25(2) : 225-251.

McOrmond-Arenja S. (2020). *'It Don't* Be *Like That No More' : Meanings and Function of Invariant* Be *in Early Black English*. Université d'Ottawa. Mémoire de maîtrise.

Mielke J. (2013). « Ultrasound and corpus study of a change from below : Vowel rhoticity in Canadian French », *University of Pennsylvania Working Papers in Linguistics* 19(2) : article 16.

Petrik K. (2005). *Deontic Modality in Quebec English : 'Everything You Need to Know'*. Université d'Ottawa. Mémoire de maîtrise.

Poplack S. (1985). « Contrasting patterns of code-switching in two communities », dans Warkentyne H. J. (éd.), *Methods V : Papers from the V International Conference on Methods in Dialectology*. Victoria, C.-B. : University of Victoria, 363-385.

Poplack S. (1989). « The care and handling of a mega-corpus », dans Fasold R. & Schiffrin D. (éd.), *Language Change and Variation*. Amsterdam : Benjamins, 411-451.

Poplack S. (éd.) (2000). *The English History of African American English*. Oxford : Blackwell Publishers.

Poplack S. (2007). « Foreword », dans Beal J., Corrigan K. & Moisl H. (éd.), *Creating and Digitizing Language Corpora*. Houndmills : Palgrave-Macmillan UK, ix-xiii.

Poplack S. (2008). « Quebec English », *Anglistik International Journal of English Studies* 19(2) (Special issue : Focus on Canadian English) : 189-200.

Poplack S. (2011). « Grammaticalization and linguistic variation », dans Heine B. & Narrog H. (éd.), *Handbook of Grammaticalization*. Oxford : Oxford University Press, 209-224.

Poplack S. (2015). « Norme prescriptive, norme communautaire et variation diaphasique », dans Kragh K. & Lindschouw J. (éd.), *Les variations diasystématiques dans les langues romanes et leurs interdépendances*, Série TraLiRo. Strasbourg : Société de linguistique romane, 293-319.

Poplack S. (2018). *Borrowing : Loanwords in the Speech Community and in the Grammar*. Oxford : Oxford University Press.

Poplack S. & Dion N. (2007). « Linguistic mythbusting : The role of the media in diffusing change », *Colloque « NWAV 36 »*, University of Pennsylvania, 11-14 octobre 2007, Philadelphie.

Poplack S. & Dion N. (2009). « Prescription vs. praxis : The evolution of future temporal reference in French », *Language* 85(3) : 557-587.

Poplack S., Dion N. & Zentz L. (2019). « L'anglicisme syntaxique : produit inévitable du contact des langues ? », *Circula : revue d'idéologies linguistiques* 9 : 78-105.

Poplack S., Jarmasz L.G., Dion N. & Rosen N. (2015). « Searching for 'Standard French' : The construction and mining of the *Recueil historique des grammaires du français* », *Journal of Historical Sociolinguistics* 1(1) : 13-56.

Poplack S. & Lealess A.V. (2009). « Language change over the lifespan revisited : Further insights from the 'Up' series », *Colloque « NWAV 38 »*, Université d'Ottawa, 22-25 octobre 2009, Ottawa.

Poplack S., Lealess, A.V. & Dion N. (2013). « The evolving grammar of the French subjunctive », *Probus* 25(1) : 139-193.

Poplack S. & Levey S. (2011). « Variabilité et changement dans les grammaires en contact », dans Martineau F. & Nadasdi T. (éd.), Le français en contact : hommages à Raymond Mougeon, *collection « Les Voies du français »*. Québec : Presses de l'Université Laval, 247-280.

Poplack S., Robillard S., Dion N. & Paolillo J.C. (2020). « Revisiting phonetic integration in bilingual borrowing », *Language* 96(1) : 126-159.

Poplack S. & Sankoff D. (1987). « The Philadelphia story in the Spanish Caribbean », *American Speech* 62(4) : 291-314.

Poplack S. & St-Amand A. (2009). « Les *Récits du français québécois d'autrefois* : reflet du parler vernaculaire du XIX^{e} siècle », *Revue canadienne de linguistique* 54(3) : 511-546.

Poplack S. & Tagliamonte S. (1989). « There's no tense like the present : Verbal *-s* inflection in Early Black English », *Language Variation and Change* 1(1) : 47-84.

Poplack S. & Tagliamonte S. (1991). « African American English in the diaspora : Evidence from old-line Nova Scotians », *Language Variation and Change* 3(3) : 301-339.

Poplack S. & Tagliamonte S. (1996). « Nothing in context : Variation, grammaticization and past time marking in Nigerian Pidgin English », dans Baker P. & Syea A. (éd.), *Changing Meanings, Changing Functions. Papers Relating to Grammaticalization in Contact Languages*. Westminster, UK : University Press, 71-94.

Poplack S. & Tagliamonte S. (2000). « The grammaticization of *going to* in (African American) English », *Language Variation and Change* 11(3) : 315-342.

Poplack S. & Tagliamonte S. (2001). *African American English in the Diaspora*. Oxford : Basil Blackwell.

Poplack S. & Tagliamonte S. (2010). « African Nova Scotian English in an enclave », dans Gold E. & McAlpine J. (éd.), *Canadian English : A Linguistic Reader*. Kingston : Strathy Language Unit, Queen's University, 146-154.

Poplack S., Torres Cacoullos R., Dion N., de Andrade Berlinck R., Digesto S., LaCasse D. & Steuck J. (2018). « Trajectories of change in Romance sociolinguistics », dans Ayres-Bennett W. & Carruthers J. (éd.), *Manual of Romance Sociolinguistics*. Berlin/Boston : de Gruyter, 217-252.

Poplack S., Van Herk G. & Harvie D. (2002). « 'Deformed in the dialects' : An alternative history of non-standard English », dans Trudgill P. & Watts D. (éd.), *Alternative Histories of English*, 87-110. London : Routledge.

Poplack S., Walker J.A. & Malcolmson R. (2006). « An English 'like no other' ? : Language contact and change in Quebec », *Revue Canadienne de linguistique* 51(2/3) : 185-213.

Poplack S., Zentz L. & Dion N. (2012). « Phrase-final prepositions in Quebec French : An empirical study of contact, code-switching and resistance to convergence », *Bilingualism : Language and Cognition* 15(2) : 203-225.

Sankoff D., Poplack S. & Vanniarajan S. (1990). « The case of the nonce loan in Tamil », *Language Variation and Change* 2(1) : 71-101.

St-Amand A. (2002). *Le subjonctif suivant une expression non-verbale*. Université d'Ottawa. Mémoire de maîtrise.

Tagliamonte S. (1991). *A Matter of Time : Past Temporal Reference Verbal Structures in Samaná English and the Ex-Slave Recordings*. Université d'Ottawa. Thèse doctorale.

Tagliamonte S., Poplack S. & Eze E. (1997). « Plural marking patterns in Nigerian Pidgin English », *Journal of Pidgin and Creole Languages* 12(1) : 103-129.

Thomason S. (2001). *Language Contact : An Introduction*. Edinburgh : Edinburgh University Press.

Torres Cacoullos R. & Walker J.A. (2009). « The present of the English future : Grammatical variation and collocations in discourse », *Language* 85(2) : 321-54.

Toth C. (2014). *Deciphering the Dative Alternation : Assessing Aspects Often Overlooked*. Université d'Ottawa. Mémoire de maîtrise.

Tottie G. & Harvie D. (2000). « It's all relative : Relativization strategies in early African American English », dans Poplack S. (éd.), *The English History of African American English*. Oxford : Blackwell Publishers, 198-230.

Van Herk G. (2000). « The question question : Auxiliary inversion in early African American English », dans Poplack S. (éd.), *The English History of African American English*. Oxford : Blackwell Publishers, 175-197.

Van Herk G. (2002). *Message from the Past : Past Temporal Reference in Early African American Letters*. Université d'Ottawa. Thèse doctorale.

Van Herk G. & Poplack S. (2003). « Rewriting the past : Bare verbs in the *Ottawa Repository of Early African American Correspondence* », *Journal of Pidgin and Creole Languages* 18(2) : 231-266.

Walker J.A. (2000). *Present Accounted For : Prosody and Aspect in Early African American English.* Université d'Ottawa. Thèse doctorale.

Walker J.A. (2001). « Using the past to explain the present : Tense and temporal reference in Early African American English », *Language Variation and Change* 13(1) : 1-35.

Willis L. (2000). '*Être ou ne plus être*' : *Auxiliary Alternation in Ottawa-Hull French*. Université d'Ottawa. Thèse de maîtrise.

Winford D. (2003). *An Introduction to Contact Linguistics*. Malden, MA : Blackwell.

Yoshizumi Y. (2006). '*She's Got an English Thing There*' *: The Variation of the Stative Possessives in Quebec City English.* Université d'Ottawa. Mémoire de maîtrise.

NOTES

***.** Les travaux dont il est question ici ont été généreusement subventionnés par le Conseil de recherches en sciences humaines du Canada par l'entremise de son programme des Chaires de recherche du Canada et de nombreuses subventions de recherche, ainsi que par la Fondation Killam, la Fondation Pierre Elliott Trudeau, le ministère de la Recherche et de l'innovation de

l'Ontario, la Fondation canadienne pour l'innovation et les Fonds ontariens pour l'innovation. C'est Bill Labov qui m'a initiée au concept de « corpus ». Le respect des données et des locuteurs qui les fournissent a toujours été au cœur de sa démarche. Les connaissances que j'ai acquises dans son célèbre cours LING 560 à l'Université de Pennsylvanie sous-tendent toutes les pratiques de collecte et de traitement des données du Laboratoire de sociolinguistique de l'Université d'Ottawa, ainsi que les cours de Dialectologie urbaine que nous donnons depuis lors. Mes efforts dans ce domaine ont été immensément secondés, puis surpassés, par des générations d'étudiants et associés brillants, engagés, enthousiastes et, surtout, extrêmement bien organisés ! Ils ont grandement contribué à traduire les enseignements de Labov en méthodes chaque fois plus performantes et efficaces. Si je peux me vanter du fait que nous parvenons à reproduire une analyse des décennies plus tard, c'est entièrement grâce à eux. Je remercie Véronique Lessard et Nathalie Dion pour leur aide précieuse avec la formulation française de cet article.

1. Anglais/igbo, anglais/tamoul, anglais/ukrainien, anglais/finnois, français/wolof, français/fongbe, français/arabe tunisien, français/vietnamien/anglais.

2. (Pour n'en nommer que quelques-uns des plus récents, voir Barysevich (2012), Kastronic (2016), Lealess (2014), Levey *et al.* (2013), McOrmond-Arenja (2020), Poplack (2018), Poplack *et al.* (2019), Poplack *et al.* (2015, 2018, 2020) et Toth (2014)). Pour des références à d'autres publications qui utilisent les corpus du Labo, consulter http://www.sociolinguistique.uottawa.ca/publications.html.

3. Il s'agit de trois sous-corpus : *Samaná English Corpus* (21 locuteurs, 22 heures d'enregistrements ; Poplack et Sankoff, 1987), *African Nova-Scotian English Corpus* (79 locuteurs, 181 heures d'enregistrements ; Poplack et Tagliamonte, 1991), *Ex-Slave Recordings* (11 locuteurs; Bailey *et al.*, 1991 ; Poplack et Tagliamonte, 1989).

4. *Corpus du français parlé à Ottawa-Hull* (120 locuteurs ; Poplack, 1989) ; *Le français en contexte : milieux scolaire et social* (166 locuteurs ; Poplack, 2015) ; *Quebec English Corpus* (183 locuteurs ; Poplack *et al.*, 2006).

5. Les noms réels sont conservés dans un endroit sécurisé et confidentiel pendant la construction du corpus, et sont détruits une fois l'anonymisation terminée.

6. Les codes entre parenthèses renvoient au nom du corpus, au numéro du locuteur et à l'adresse de l'extrait dans le *Quebec English Corpus* (Poplack *et al.*, 2006) en (1) et dans le *African Nova Scotian English Corpus* (Poplack et Tagliamonte, 1991) en (2). Les exemples sont cités verbatim à partir des enregistrements.

7. Les extraits en (1) et (2) fournissent des exemples de tels renvois.

RÉSUMÉS

Cet article détaille les principes et pratiques de gestion de données en vigueur au Laboratoire de sociolinguistique de l'Université d'Ottawa (le Labo ; http://www.sociolinguistique.uottawa.ca/laboratoire.html), centre qui abrite 19 importants corpus correspondant à des centaines d'heures d'enregistrement de parler spontané. Notre propos s'inscrit dans le cadre de la sociolinguistique variationniste et fournit un aperçu des méthodes éprouvées en matière de constitution de corpus, qui comprend notamment la collecte, la transcription, l'annotation, le repérage, le codage, et l'analyse des données. Nous abordons également la préservation et le cycle de vie des données, et jetons un coup d'œil aux considérations déontologiques qui caractérisent la collecte

et l'analyse du vernaculaire. Nous concluons par un survol des nombreuses applications linguistiques possibles des données de parler spontané bien gérées.

This article details the data management principles and practices developed by the University of Ottawa Sociolinguistics Lab (http://www.sociolinguistics.uottawa.ca/thelab.html), home to 19 major corpora representing hundreds of hours and millions of words of recorded everyday speech. Couched within the variationist framework for linguistic analysis, it provides a practical overview of tried-and-true methods for corpus construction, including data collection, transcription, annotation, and citation, as well as data retrieval, coding, and analysis. It also features observations on data preservation and data lifecycle, and discusses ethical considerations involved in collecting and analyzing vernacular speech. It concludes with a summary of the wide variety of linguistic applications to which properly managed spontaneous speech data can be put.

INDEX

Keywords : Data management, Data collection, Corpus construction, Data transcription, Speech data, Variationist sociolinguistics
Mots-clés : Gestion de données, Collecte de données, Constitution de corpus, Transcription de données, Données de production orale, Sociolinguistique variationniste

AUTEUR

SHANA POPLACK

Université d'Ottawa