Introduction

Céline Dugua, Layal Kanaan-Caillol

CORPUS

Corpus

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

Introduction

Céline Dugua et Layal Kanaan-Caillol

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/5885
DOI : 10.4000/corpus.5885
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Céline Dugua et Layal Kanaan-Caillol, « Introduction », *Corpus* [En ligne], 22 | 2021, mis en ligne le 12 février 2021, consulté le 16 février 2021. URL : http://journals.openedition.org/corpus/5885 ; DOI : https://doi.org/10.4000/corpus.5885

Introduction

Céline Dugua et Layal Kanaan-Caillol

1 Depuis les années soixante-dix et le corpus de Montréal (Sankoff *et al.*, 1976), les corpus oraux et multimodaux ont été au cœur des transformations technologiques, méthodologiques et théoriques de la linguistique sur corpus numériques, reconfigurant les attentes en matière de conservation des documents sonores. Les outils et instruments de transcription, d'annotation, de traitement du signal, de textométrie, de visualisation, et plus généralement tous les outils du TAL et du traitement de données, les plateformes de conservation et de diffusion de corpus, les initiatives visant l'interopérabilité des données sont apparus comme indissociables des analyses et des opérations de constitution et d'exploitation de corpus.

2 Enfin, à l'heure du web de données, les questions posées par l'archivage et par la réutilisation de corpus, tout comme les projets de sciences contributives dépassent le domaine de la linguistique bien que celle-ci soit concernée en premier lieu. Ces transformations, qui nécessitent une réflexion sur la normalisation et le formatage, questionnent la place qui doit être faite à des données hétérogènes pour l'étude de la variation.

3 C'est autour de ces problématiques des corpus oraux qu'avec des collègues du LLL-UMR7270, nous avons organisé en novembre 2018, le colloque anniversaire des 50 ans des Enquêtes Sociolinguistiques à Orléans (ESLO). Intitulé « 50 ans de linguistique sur corpus oraux : apports à l'étude de la variation », ce colloque a accueilli une large communauté de linguistes, de sociolinguistes, de Talistes présentant leurs travaux sur des corpus francophones (France, Québec) et italien, japonais, anglo-américain, créole haïtien. Au-delà de cet évènement, nous avons souhaité, dans ce numéro de *Corpus*, élargir le panorama en rassemblant des travaux qui problématisent ce qu'impliquent la prise en compte et l'étude des variations sur les outils utilisés à chaque étape de la chaine de traitement d'un corpus, en commençant par la constitution des données. Cette chaine de traitement a été formalisée pour ESLO par Baude (2006) et Baude et Dugua (2011, 2016) qui soulignent par ailleurs les liens constants et multidirectionnels entre les différentes phases du process. Avec une approche sociolinguistique variationniste, ESLO s'est proposé de constituer un corpus prototypique qui puisse

mettre en évidence les conséquences que la prise en compte de la variation a sur l'ensemble des étapes de la chaine.

4 La question centrale posée dans ce numéro est justement cette prise en compte de la variation, ou plutôt de toutes les variations, y compris celles inhérentes aux tâches de transcription ou de catalogage, dans le recueil et la constitution de corpus ainsi que dans les outils de traitement et d'analyse. Les huit articles sélectionnés recensent les contraintes et les solutions envisagées pour répondre à cet objectif, et l'un d'eux est consacré à la question cruciale de l'impact des nouvelles réglementations sur la collecte des données.

5 Nous distinguons quatre thématiques dans ces contributions : la constitution de corpus, les métadonnées, l'outillage et les objets variables. Le volet éthique et juridique constituera un cinquième point.

1. Constitution de corpus

6 Dans la chaine de traitement, chaque étape est conditionnée par les précédentes et anticipe les suivantes, en intégrant la variation, comme l'article de Sh. Poplack qui ouvre ce numéro en apporte la démonstration. L'auteure présente différents corpus du Labo (Laboratoire de linguistique de l'Université d'Ottawa) en mettant l'accent sur plusieurs phases, de la collecte, à la conservation en passant par la transcription, l'annotation, l'analyse.

7 L'étape de la collecte détermine l'ensemble du processus et conditionne la phase d'exploitation. De ce fait, la constitution de corpus, notamment dans sa collecte répond, ou en tout cas est guidée, par des questions liées à l'objet/aux objets d'étude. C'est dans ce sens que Sh. Poplack affirme :

> 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. (Poplack, ce numéro).

8 « Quoi recueillir et auprès de qui ? ». Plusieurs articles de ce numéro reprennent cette question en fonction des visées linguistiques propres à chaque projet. Il en résulte une large palette de choix méthodologiques autour d'un enjeu central : appréhender la variation.

9 Saisir la variation dès l'étape de la collecte commence par la définition des caractéristiques attendues des participants :

- corpus stratifié pour Blondeau *et al.*, comparable avec la première enquête sociolinguistique sur le français de Montréal (1971) pour observer le changement en temps réel afin de vérifier l'hypothèse selon laquelle la variation linguistique et sociale synchronique est à la base de la variation diachronique (Labov, 1994).
- enfants issus de familles de milieux socio-économiques contrastés, « à l'image de ce que peut être la population d'une ville » pour J. Ganaye qui se propose de « questionner l'influence des milieux socio-économiques sur l'usage du langage » dans le cadre de l'étude de la liaison.
- élèves et adultes dans une école maternelle mixte pour A. Nardy *et al.* en vue d'une étude de la variation et des réseaux sociaux sur des données massives.

10 Saisir la variation c'est aussi recueillir des données authentiques, spontanées ; recueillir le *vernaculaire*. Cela implique, pour le chercheur, de créer des situations de collecte favorables et de résoudre le *Paradoxe de l'observateur* (Labov, 1972). L'*entrevue sociolinguistique* (Labov, 1984) constitue un des dispositifs le plus souvent adopté.

11 Le recueil des données du corpus Hochelaga-Maisonneuve (Blondeau *et al.*) s'est fait dans le cadre d'entretiens en face-à-face au domicile des participants, menés pour certains par les chercheurs du projet et pour d'autres par des étudiants. Un guide d'entretien a servi de support tout en gardant l'objectif d'un échange fluide et spontané autour de sujets tels que les souvenirs d'enfance, les événements biographiques, la vie de quartier et des questions liées à l'usage du français à Montréal et au Québec.

12 Pour certains corpus du Labo (Sh. Poplack), les chercheurs ont formé des membres des communautés ciblées afin de leur confier la réalisation des entretiens.

13 J. Ganaye pour sa part a fait le choix de l'absence du chercheur. Pour étudier l'acquisition du langage chez les enfants – à travers l'étude de la liaison – en prenant en compte l'impact de l'input (l'environnement langagier) sur l'output (productions enfantines) dans différents environnements, le corpus a été réalisé par saisie « [de] situations naturelles variées formant le quotidien des enfants ». Avec un kit d'enregistrement et un mode opératoire confiés aux parents, l'absence du chercheur vise à « réduire l'effet du « paradoxe de l'observateur » et [à] accéder aux situations les plus écologiques possibles », telles que les repas, les moments de jeux, les devoirs, etc.

14 A des fins de recueil de données massives et longitudinales pour l'étude des dynamiques langagières des élèves en lien avec les réseaux sociaux, et ce dans une école maternelle socialement mixte, A. Nardy *et al.*, quant à eux, ont recours à un dispositif de collecte embarqué, très spécifique : « Environ 200 individus (enfants et adultes) sont équipés une semaine par mois pendant 3 ans de capteurs qui enregistrent en continu à la fois leurs interactions verbales et leurs contacts sociaux. »

15 En somme, ces quatre exemples illustrent la variété des protocoles de collecte, depuis des démarches classiques et largement répandues qui font leur preuve, jusqu'à des procédures spécifiques, déterminées par l'objet de recherche et les relations entre les locuteurs.

2. Métadonnées

16 Dans la constitution d'un corpus variationniste, les métadonnées revêtent une importance particulière. La nomenclature et le recueil des métadonnées doivent être pensés en amont de la collecte, comme le mentionnent H. Blondeau *et al.* et J. Ganaye. Sur des échantillons différents et avec des objectifs d'analyse spécifiques, les types de métadonnées ne seront pas identiques. J. Ganaye (ESLO-Enfants) privilégie des informations concernant les familles (types d'activités pratiquées, rapport à la culture, CSP, etc.) et les résultats à des tests de langage que les enfants ont passés. Dans le cadre du corpus FRAN-HOMA (Blondeau *et al.*) les auteures insistent sur l'importance de la stratification et des variables nécessaires à son organisation.

17 Abordant les corpus de parole pathologique en français, avec l'objectif de les rassembler en base de données, A. Ghio *et al.* soulignent la nécessité du lien entre les données sonores, les données transcrites et les métadonnées – notamment les caractéristiques cliniques des locuteurs. Cet enjeu se retrouve dans ESLO et dans FRAN-

HOMA pour les métadonnées sociodémographiques et la description des situations de collecte. L'accès à des métadonnées (riches) permettra de travailler sur des données situées et d'éclairer le travail sur des variables linguistiques. Dans ces deux corpus, qui intègrent la dimension microdiachronique, ces informations sur les locuteurs participent à la construction d'une comparabilité entre des données recueillies à des périodes différentes, et permettent de mieux comprendre les dynamiques de changement linguistique en temps réel et en temps apparent.

18 A. Ghio *et al.* soulignent que de nombreux obstacles ont été levés concernant les métadonnées mais que « le maillon faible reste la normalisation et la structuration des données sur les locuteurs et leurs productions langagières ». Le cas de la parole pathologique interroge sur la finesse des informations cliniques des participants et sur les aspects juridiques qui protègent ces données, comme cela est évoqué par M. Lalain *et al.*

19 Le lien entre données et métadonnées est également au centre de l'article de Fl. Badin *et al.* qui mettent en place une méthodologie de traitement à partir d'une collection du corpus ESLO dans lequel les métadonnées sont structurées en respectant le *Dublin-Core* avec des champs sociolinguistiques plus riches. Le travail réalisé par ces auteurs permet, en conservant toute la richesse des métadonnées du locuteur, d'interroger l'ensemble du corpus ESLO dans TXM, en conservant un accès permanent aux métadonnées.

3. Outillage, de la collecte à l'analyse et à la mise à disposition

20 Aujourd'hui, la question de l'outillage est présente à chaque étape de la constitution du corpus. Dans le recueil de parole au sein de la famille et autour des enfants, J. Ganaye a privilégié des captations doubles audio et vidéo qui se devaient d'être simples à utiliser car le matériel était installé par les participants eux-mêmes.

21 Avec des collectes dans toute une école maternelle, sur plusieurs années, croisant données langagières et données de réseaux, A. Nardy *et al.* ont mis en place un dispositif d'enregistrement audio embarqué original qui intègre dans un boitier de petite taille porté par les participants (adultes et enfants) à la fois un micro, un dispositif d'horodatage et une capacité de stockage. Une chaine de traitement outillée est ensuite nécessaire pour traiter la masse de données recueillies. Au niveau des signaux, un traitement automatique permet de déterminer par la voix quel enfant portait le boitier (dans une classe de maternelle, c'est une information difficile à obtenir) et d'en faire un pré-traitement automatique qui facilitera la transcription, notamment en bornant les phases de prise de parole et de silence.

22 Les articles de ce numéro font référence à différents outils de transcription et d'annotation (*Transcriber*, *Elan*, *Praat*, *Clan*) qui, comme le soulignent H. Blondeau *et al.* « provide greater flexibility and precision in the analysis of the dynamics of sociolinguistic variation. ». Par ailleurs, plusieurs pré-traitements de la transcription pour l'annotation sont présentés notamment pour le repérage des variables linguistiques (Wu et Adda-Decker, Nardy *et al.*) et pour l'annotation morpho-syntaxique (Badin *et al.*).

23 Un pré-traitement des transcriptions pour faciliter l'alignement au signal est exposé par A. Nardy *et al.* ; le caractère chronophage de la segmentation avait été mis en évidence par Baude et Dugua (2011). Dans la méthode utilisée par A. Nardy *et al.*, les segments sont pré-indiqués et le transcripteur n'a plus qu'à ajuster les frontières, si nécessaire. Les auteurs signalent également un ensemble de scripts qui fournissent autant d'outils aux transcripteurs pour gagner en temps et en qualité, en particulier pour la vérification des annotations phonologiques et l'anonymisation.

24 Y. Wu et M. Adda-Decker ont fondé leur étude sur le corpus *Nijmegen Corpus of Casual French* (NCCFr) (Torreira et Ernestus, 2010) en appliquant une méthode d'alignement forcé entre les transcriptions existantes et le signal de parole. L'originalité de leur approche réside dans la recherche de phénomènes de variation phonétique à partir d'écarts entre la prononciation prédite à partir des mots orthographiques et celle obtenue en tenant compte du signal de parole à partir du système de reconnaissance automatique de la parole du LIMSI (Gauvain *et al.*, 2005).

25 L'outil TXM utilisé par Fl. Badin *et al.* est un « outil de textométrie puissant, capable de gérer une annotation morphosyntaxique ainsi que la richesse des métadonnées du corpus » (Badin *et al.* ce numéro). Les auteurs exposent la chaine de traitement mise en place pour créer une version TXM du corpus ESLO qui permette une visualisation du corpus explicite (prenant en compte la question des unités de segmentation), une partition en sous-corpus en conservant l'accès aux métadonnées et une interrogation à la fois sur les occurrences, les lemmes et les catégories morphosyntaxiques. Les évolutions récentes de TXM, au départ conçu pour les corpus écrits, permettent au chercheur d'accéder au son à tout moment et constitue une boucle prometteuse pour l'analyse du corpus ESLO : des annotations depuis les transcriptions, la possibilité de partitionner un corpus, en ayant toujours accès 1) aux métadonnées des enregistrements et des locuteurs et 2) aux enregistrements.

26 Si la masse des données traitées dans les grands corpus oraux nécessite le recours aux outils, il faut se garder de la distance qu'ils introduisent entre le chercheur et ses données. Nous adhérons, en ce sens, au point de vue de Sh. Poplack (ce numéro) :

> 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. Objets variables et Dia-Variations

27 Bien que la thématique de ce numéro repose de manière centrale sur les aspects méthodologiques de la constitution de corpus « du recueil à l'outillage », plusieurs articles ont illustré la manière dont les choix et les dispositifs méthodologiques donnent accès à la variation à travers la présentation d'études sur des objets variables. Ils mettent en lumière la relation consubstantielle entre données et métadonnées dès lors qu'il s'agit de saisir la variation.

28 Avec une visée microdiachronique, la constitution de nouveaux corpus répond au besoin de créer des fenêtres temporelles permettant l'observation du changement à la manière de l'enquête ESLO2 du corpus ESLO.

29 C'est le cas aussi du corpus FRAN-HOMA, présenté par H. Blondeau *et al.*, construit en rapport avec le corpus de Montréal de 1971 pour observer le changement à 40 ans d'intervalle. Pour H. Blondeau *et al.*, un des objectifs majeurs est de questionner les modèles théoriques, notamment celui du temps apparent « afin de vérifier ses prédictions sur les observations en temps réel » (Blondeau *et al.*, ce numéro). Les auteures soulignent que le nouveau corpus, fort des avancées technologiques mobilisées tout au long de la chaine de traitement, rend possibles de nouvelles recherches sur la variation. Elles en donnent une illustration en présentant les résultats de l'étude de deux variables sociolinguistiques – la variation dans l'emploi des marqueurs discursifs à travers l'étude de *fait-que* dans ses variantes phonologiques / fɛk/ et /fak/ et la contraction de la préposition *dans* – dans une perspective diastratique et/ou diachronique.

30 Dans un autre champ, celui de l'acquisition du langage, en choisissant d'enregistrer différentes situations du quotidien d'enfants avec leurs parents et en contrastant les catégories socio-culturo-professionnelles des familles, J. Ganaye intègre la variation diastratique et diaphasique dont elle croise les paramètres avec l'observation du développement langagier. Il s'agit pour elle de comprendre la façon dont le langage – tout particulièrement un phénomène variable : la liaison – se construit dans la diversité des environnements et des situations auxquels se trouve confronté un enfant.

31 En intégrant approches micro-diachronique, diastratique et diaphasique, Fl. Badin *et al.*, illustrent l'intérêt de leur méthodologie à partir de l'étude de l'emploi des interrogatives partielles. Ils relèvent un changement en faveur des interrogatives partielles *in situ* (ex. *Tu pars quand ?* vs. *Quand tu pars/Quand pars-tu ?*) à travers un jeu de requêtes guidé par les résultats successifs combinant angle diastratique et angle diaphasique.

32 Les objets variables peuvent également être saisis en termes de variation interne. Y. Wu et M. Adda-Decker s'intéressent aux prononciations, à travers l'étude des phénomènes de réduction en parole continue. Avec un outillage calibré et des dictionnaires de prononciation en référence pour les réalisations canoniques, les auteures saisissent des variations paradigmatiques et syntagmatiques et identifient les segments les plus accessibles à la réduction en intégrant les caractéristiques intrinsèques des sons et les effets phonotactiques.

5. Volet éthique et juridique

33 Nous terminerons cette introduction par un aspect qui revêt une importance particulière dans la constitution, le traitement, la diffusion des corpus oraux : celui du cadre éthique et juridique.

34 La recherche des bonnes pratiques est au centre des préoccupations des linguistiques de corpus (Baude *et al.*, 2006). Le recueil du consentement éclairé des locuteurs ou l'anonymisation des données (le remplacement des noms des locuteurs par des codes, l'anonymisation dans les transcriptions et dans les enregistrements), qui font partie des pratiques depuis plusieurs décennies sont abordées à plusieurs reprises dans ce numéro.

35 C'est en termes de « considérations d'ordre déontologiques » que Sh. Poplack présente le recueil du consentement éclairé, qu'il se fasse avant ou après la collecte, et les

différents mécanismes « [assurant] la confidentialité des données » : l'anonymisation de l'identité des participants, la sécurisation des données et leur consultation. La section « Formulaire de consentement, aspects juridiques et éthiques » de l'article de J. Ganaye précise le contenu du formulaire de consentement proposé aux familles participantes. A. Nardy *et al.* évoquent les procédures de validation par différentes instances (COERLE de l'INRIA, CNIL). Enfin, A. Ghio *et al.* soulignent que le volet juridique ne concerne pas uniquement la constitution de corpus mais aussi la récupération de corpus.

36 L'article de M. Lalain *et al.* a un statut particulier par rapport à la thématique du fait qu'il aborde essentiellement les questions éthiques et juridiques. Il sera d'une grande aide pour les chercheurs qui travaillent sur corpus puisqu'il permet de préciser les contours du RGPD et de la loi Jardé, les difficultés et les solutions répondant à leurs injonctions. Il offre des pistes sur la façon dont nous, chercheurs, pouvons et devons nous l'approprier afin de travailler conformément au cadre légal et réglementaire. Les auteurs concluent leur article en montrant que ces nouvelles réglementations peuvent être vues de manière positive puisqu'elles amènent le chercheur et la recherche en général à mieux protéger les personnes.

Conclusion

37 À travers les observations synthétiques autour des cinq axes que nous avons privilégiés, nous voyons se dessiner la cohérence de travaux issus de champs différents avec des objectifs scientifiques hétérogènes. L'aperçu donné dans cette introduction est une invitation à la découverte des articles.

38 La constitution de corpus est une entreprise d'une grande envergure qui nécessite des moyens humains, techniques et financiers conséquents et les réponses en termes de financements ne sont souvent pas à la hauteur des enjeux liés d'une part à la patrimonialisation et d'autre part à l'étude des dynamiques langagières.

39 Nous rejoignons Sh. Poplack lorsqu'elle écrit :

> 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. (Poplack, ce numéro)

40 Ce que nous souhaitons retenir en conclusion et qui, en réalité, est démontré dans chacun des articles, ne serait-ce qu'en filigrane, est le fait que la recherche linguistique commence dès le premier maillon de la chaine de traitement. La constitution de corpus se fait nécessairement à la lumière des questions scientifiques et chacun des choix, à chacune des étapes, est sous-tendu par un ancrage théorique qui guide l'ensemble des opérations et des analyses subséquentes.

BIBLIOGRAPHIE

Baude O. (2006). *Corpus oraux : guide des bonnes pratiques.* CNRS-Editions et Presses universitaires d'Orléans.

Baude O. & Dugua C. (2011). « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? ». *Corpus* 10 : 99-118.

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

Bergounioux G., Jacobson M. & Pietrandrea P. (2017). « L'annotation des corpus oraux », in Ayres-Benett W. & Carruthers J. (éd.) *Manual of Romance Sociolinguistics.* Berlin, De Gruyter : 27-58.

Habert B., Nazarenko A. & Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Labov W. (1972). *Language in the Inner City : Studies in the Black English Vernacular*. Philadelphie : University of Pennsylvania Press.

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

Labov W. (1994). *Principles of Linguistic Change. Volume I : Internal Factors.* Oxford and Malden : Blackwell.

Ochs E. (1979). « Transcription as theory », in E. Ochs & B. Schieffelin (éd.) *Developmental pragmatics*. New York : Academic Press, 43-72.

Sankoff D., Sankoff G., Laberge S. & Topham M. (1976). « Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale », Cahiers de linguistique 6 : 85-125.

AUTEURS

CÉLINE DUGUA

Laboratoire Ligérien de Linguistique (LLL-UMR7270), Université d'Orléans

LAYAL KANAAN-CAILLOL

Laboratoire Ligérien de Linguistique (LLL-UMR7270), Université d'Orléans