Comparer des données de corpus : évidence, illusion ou construction ?

Françoise Gadet, Sandrine Wachs

▶ **To cite this version:**

Françoise Gadet, Sandrine Wachs. Comparer des données de corpus : évidence, illusion ou construction ?. Langage et Société, 2015. hal-01282742v2

HAL Id: hal-01282742

https://univ-sorbonne-nouvelle.hal.science/hal-01282742v2

Submitted on 4 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Gadet F. & Wachs S. (2015), « Comparer des données de corpus : évidence, illusion ou construction ? » Langage et Société 154, Maison des Sciences de L'homme Paris, 33-49.

Comparer des données de corpus : évidence, illusion ou construction ?

Françoise Gadet (UPO & MoDyCo) et Sandrine Wachs (Sorbonne Nouvelle & DILTEC)

Cet article étudie la propriété de *comparabilité*, trop souvent prise comme une évidence. Il s'arrête à la relation entre recueil de (grands) corpus et travail de terrain, quant à ce que les différents gestes induisent comme questionnements sur la langue et sa variabilité. Après avoir rappelé les démarches de grands corpus de français, il s'arrête au cas d'un corpus récent recueilli en région parisienne, MPF, pour finir par montrer quelles grandes questions sociolinguistiques pourraient être ouvertes à partir de la comparabilité.
Mots-clés : Corpus, terrain, variation, variabilité, comparabilité, français.

This article studies the property of *comparability*, which is too often taken as self-evident. It investigates the relationship between the collection of (large) corpora and fieldwork, and the questions induced by each about language and language variability. After summarizing some approaches adopted by large corpora of French, one particular corpus recently collected in the Paris region, the MPF, is discussed in greater detail. The study concludes by showing what broad sociolinguistic questions may be raised through the notion of comparability.
Keywords: Corpora, fieldwork, variation, variability, comparability, French.

Introduction[1]

Une bonne part du travail en (socio)linguistique consistant à *comparer* (des phénomènes, des structures, des énoncés, des événements discursifs et, pour ce dont on traitera ici, des corpus ou des parties de corpus), la *comparabilité* mérite d'être considérée en soi, au-delà des évidences. La question est d'autant plus vive que les « (grands) corpus » sont désormais omniprésents en sciences du langage, avec justement la comparabilité parmi leurs objectifs. Le sociolinguiste a dès lors à se demander si la pratique du *terrain* est compatible avec le recueil d'une importante masse de données, supposées exploitables par d'autres que ceux qui les ont recueillies.

Cet article prendra appui sur des corpus de français du point de vue de la comparaison pour faire émerger des questions sociolinguistiques sous-jacentes aux gestes de construction d'observables (ce qu'ils supposent sur la langue, sur le social, sur leur relation et sur la variation). On commencera par évoquer des corpus « historiques » pour lesquels la comparaison était d'une façon ou d'une autre à l'ordre du jour. La section 2 évoquera un corpus récemment constitué en région parisienne, où les options de recueil pourraient sembler barrer la possibilité même de comparer. On terminera en revenant aux notions de *comparaison* et de *comparabilité*, et à ce qu'elles impliquent en sociolinguistique.

[1] Merci à Paul Cappeau et Médéric Gasquet-Cyrus pour leurs relectures et leurs suggestions.

1. Comparer en supposant du comparable : histoires de corpus oraux de français

Les « grands corpus » ne se focalisent pas tous sur leurs modalités de recueil, l'attention allant en général plutôt à la masse collectée. Tous n'ont pas non plus des visées de comparaison, directes ou implicites. Quand objectif de comparaison il y a, il est en général posé dès l'amont, selon des figures que nous résumerons en « retour sur un terrain », « retour sur un terrain avec réélaboration », et « protocole conçu pour la comparaison », qu'on illustrera de quelques exemples. Il est à noter qu'il est rare, dans la littérature sur ces corpus, de lire des réflexions explicites sur le sens de ces retours, souvent pris comme un geste ordinaire.

1.1. Réplique de méthodes antérieures

Les retours sur un terrain ne sont pas une démarche fréquente dans la brève histoire des corpus de français, ce qui est manifeste dans un inventaire (Gadet 2013) qui couvre 145 corpus de français « hors de France »[2]. On peut évoquer les 3 phases des corpus de Montréal (voir Vincent 2009), les corpus de Mougeon en Ontario, de King dans les Provinces Maritimes canadiennes… Il y en a encore moins en Afrique, où peu de chercheurs ont adopté une démarche de type variationniste. Les retours ont aussi été rares pour le français de France.

Quand les corpus viennent d'un même chercheur ou d'une même équipe, faisant ainsi de la conformité à une méthode initiale une condition (il est souvent dit « garantie ») de comparabilité, le protocole est reproduit. Les paramètres de sélection des locuteurs sont alors toujours « externes » (socio-démographiques), les seuls à apparaître aisément réplicables.

La démarche est la même, en plus exigeante, dans le corpus Montréal 84 (voir Thibault et Vincent 1990) : les conceptrices du corpus ont retrouvé la moitié des locuteurs du corpus Montréal 1971. Ceux-ci ont évidemment changé d'âge, parfois de statut social, mais le fait qu'il s'agisse des mêmes locuteurs permet une réflexion sur les manifestations linguistiques des trajectoires de vie (voir par exemple Blondeau *et al*. 2002, sur les effets linguistiques de la mobilité sociale).

1.2. Réajuster, sur un même terrain, des méthodes antérieures

Revisiter un cadre théorique, tel a été par exemple l'objectif d'ESLO2 (*Etude SocioLinguistique à Orléans*), élaboré « en écho à ESLO » une quarantaine d'années après. ESLO2 reprend l'idée d'échantillonnage d'ESLO (Baude et Dugua 2011), en la revisitant à la lumière de « progrès théoriques et technologiques », concernant avant tout la transcription et la classification sociale

[2] Voir Gadet (2011) sur l'apport des corpus de français hors de France pour étudier le français en général.

des locuteurs. ESLO2 soigne les étapes de la transcription et des critères de « représentativité », en particulier l'échelonnage des locuteurs et la diversification des situations de communication, raffinant ESLO qui s'appuyait sur des catégories de l'INSEE avec l'idée de « locuteurs représentatifs ». La comparaison est souvent alléguée, mais sans élaboration autre, semble-t-il, que l'appui sur des paramètres externes. La tentative pour croiser les critères socio-économiques avec la mobilité sociale et le capital culturel est réinvestie par ESLO2. Les situations de communication, quant à elles, sont affinées selon une échelle complexe de degrés de formalité.

L'équipe canadienne de Vincent a elle aussi réévalué un cadre antérieur en ouvrant des corpus recueillis sur des bases variationnistes vers l'analyse de conversation (voir Vincent 2009), ce qui a exigé de préférer des données « naturelles » aux entretiens.

À cette rareté des retours, on peut supposer deux causes opposées : la reconnaissance de ce que la réplique des conditions de recueil n'est qu'illusion, donc inutile (voir plus loin), ou l'absence de réflexion sur les méthodes, que la plupart des grands corpus n'investissent pas beaucoup.

1.3. La comparabilité mise à la base d'un projet

Prévoir en amont, c'est encore davantage le cas, pour des raisons évidentes, dans les grands corpus multi-sites. Ainsi *CIEL_F* recherche « les mêmes » événements écologiques dans différentes situations de francophonie : il a d'abord fallu établir quels événements discursifs pourraient être regardés comme suffisamment proches d'une aire à l'autre pour apparaître *comparables*, à défaut de *similaires* (Gadet *et al*. 2012[3]). Pour *PFC*, la sélection concerne aussi les profils des locuteurs, à travers une « méthodologie commune » (Durand *et al*. 2009[4]) : les corpus, recueillis dans de nombreuses aires de la francophonie, doivent permettre de comparer des phénomènes phoniques sur la base d'une méthodologie reconduite dans chaque site. Quant au *CRFP* (DELIC 2004), ses fondements sont plutôt dans de grandes catégories de « genres » (comme *privé*, *public*, *professionnel*).

Ces différents corpus suivent différentes modalités de recueil, selon un protocole fixe pour *PFC*, avec une latitude laissée aux équipes locales pour *CIEL* – recueil écologique oblige, évidemment impossible à pré-formater. Dans tous les cas, il est supposé que la comparabilité

[3] « La comparabilité est fondée sur une sélection de situations de communication, qui, à un certain niveau d'abstraction, sont observables dans un grand nombre de contextes socioculturels actuels » (Gadet *et al.* 2012, p. 40).

[4] Un rôle dans la comparabilité est aussi prêté aux métadonnées : « Ces métadonnées assurent la traçabilité et rendent possible la comparabilité des données. » (Durand *et al*. 2009, p. 10). *Traçabilité*, certes, mais comment passe-t-on de là à *comparabilité* ?

doit/peut être assurée par un contrôle de paramètres relevant de catégories externes, des locuteurs (profil socio-démographique ou socio-culturel) et/ou des situations (aire et conditions de recueil, modalités d'entrée en contact, thèmes traités, type d'enquête, genre discursif, etc.).

1.4. *Comparer* dans la littérature sociolinguistique

Malgré le rôle prêté à la comparaison, la littérature de la discipline – manuels, ouvrages d'introduction, dictionnaires, *handbooks*, articles méthodologiques – demeure discrète sur les questions que celle-ci soulève, aussi bien méthodologiques (fondements et conditions de comparabilité) que théoriques : que s'agit-il d'assurer ? Y a-t-il du *même* en sociolinguistique (à quel prix ?) ? Et qu'en est-il de l'assignation des frontières entre du même et de l'autre ? (Nicolaï et Ploog 2013).

La sociolinguistique ne fait en général pas de la comparabilité une question cruciale. À une exception près (Tagliamonte 2002), nous n'avons pas trouvé d'entrée ou d'article « comparer »/« comparaison » et encore moins « conditions de comparabilité » dans les dictionnaires (Moreau 1997), les introductions (Meyerhoff 2006, parmi d'autres), ou les *handbooks*, français ou anglo-saxons (Coulmas 1997, Chambers *et al.* 2002, Simonin et Wharton 2013, Chambers and Schilling 2013, etc.). Et l'exception Tagliamonte n'en est pas une, portant sur la méthode comparatiste, non sur les données, dont la comparabilité est vite admise (2002, p. 733).

Ce qui se résume à une "collection of taken for granted propositions" (Cameron 1990, p. 79), avec des catégories sociales reprises du sens commun et recoupant souvent les mythes langagiers de la doxa, se montre actif chez beaucoup de linguistes élaborant des corpus, souvent sans qu'ils l'aient explicitement voulu, mais avec d'évidents impacts sur leurs méthodes et les théorisations ultérieures.

1.5. Quels positionnements théoriques se profilent derrière ces pratiques ?

Les paramètres mis en avant pour établir quelles données recueillir sont « externes » (si l'on veut garder la terminologie saussurienne), indépendants de la langue. Rien là de neuf, comme le montre Cappeau (2012) qui fait une « petite histoire des métadonnées » : une caractérisation externe est déjà pratiquée par Damourette et Pichon, qui ont collecté des données orales à défaut de faire un corpus. Emergent ainsi, sans que les concepteurs de corpus l'aient spécialement

voulu, des hypothèses sur la source de la variabilité langagière[5], assignée à des catégorisations macrosociologiques attribuées par le chercheur et supposées reproductibles.

Outre le choix des locuteurs, les recueils ont privilégié les entretiens, bien au-delà du poids social très relatif de ce genre discursif. Certes, ils sont plus faciles à recueillir que des événements écologiques, qui exigent un ajustement social, mais il y a aussi là une conformité aux conceptions de la « première vague » des études de la variation (Eckert 2012).

Tous les entretiens sont alors supposés équivalents, *ipso facto* comparables. Pourtant, quel rapport entre une interaction où deux protagonistes partagent un réseau et une autre faite avec un inconnu au hasard de rencontres dans un lieu public ou un organisme ? Même si la catégorisation des acteurs peut être la même, il est probable que les produits langagiers différeront, sans parler des contenus. Mais ce point de vue suppose de regarder les interactants autrement que comme des items destinés à remplir les cases d'une grille ou des conglomérats de catégories : comme des individus ayant une identité propre ; des agents, tous différents.

Sélectionner certains paramètres (dès lors réputés pouvoir transformer un locuteur en « locuteur représentatif ») parmi les candidats au reproductible/comparable, c'est produire, souvent sans l'avoir voulu, un « gommage » (*erasing* chez Irvine & Gal 2000) des autres (connus ou à établir), dont il resterait à montrer que, même ignorés d'un quadrillage socio-démographique, ils seraient sans effets. Or, dans les pratiques ordinaires, tout ne saurait être contrôlé, toute interaction étant située, inédite par la souplesse de paramètres souvent ténus – remarque qui ne doit pas inciter à renoncer à généraliser.

L'ordre du *sociolinguistique* semblerait alors se borner au socio-démographique, bien que toute interaction soit en constante dynamique porteuse de sens social – comme l'avait montré Labov à Marthas's Vineyard (1972)[6].

1.6. Invisibilisations et gommages

La comparabilité est ainsi appuyée sur des paramètres formalisables, quantifiables, en dichotomie (le sexe) ou par strates (l'âge ou la classe sociale). Cela revient à décréter, sur une saisie ne mettant pas la langue en jeu, que les données sont comparables, donc qu'on peut les comparer. Pourquoi ces paramètres-là ? Par routine, ou parce qu'on a établi leur signification

[5] Il n'y a pas là une critique des corpus oraux existants, dont il s'agit de passer au crible les fondements et les présupposés. On verra d'ailleurs que certaines de nos remarques s'appliquent tout autant à *MPF*. Il faut admettre qu'aucun corpus n'a que des qualités, ce qui n'empêche pas de débusquer les "hidden assumptions" (Cameron 1990, p. 79) qui reflètent le caractère d'ensemble construit de tout corpus.

[6] Parmi d'autres ayant souligné le rôle fondateur, mais aussi négligé, de ce texte, voir Eckert (2012) ou Coupland (2007).

sociolinguistique pour ce terrain-là ? La quête de diversification tourne alors à l'homogénéisation des données (« illusion d'objectivation », Le Bianic *et al.* 2012, p. 15). Quant aux métadonnées, elles renseignent aussi d'abord des catégories préconçues, alors qu'une interaction authentique ne répond jamais à un format-type, au-delà de grands genres identifiables.

Quelle(s) propriété(s) sociolinguistique(s) émerge(nt) de ces paramètres *de facto* privilégiés par le désintérêt pour ceux qui sont moins quantifiables ? La propriété d'être aisément formalisable ou quantifiable est-elle autre chose que formelle, structurelle, voire fortuite ? Elle ne garantit dans tous les cas rien de sociolinguistique, supposé ou avéré, et la question de ce qui assure une diversification des productions demeure intouchée. Chambers (2003) est parmi les rares à s'interroger, longuement (p. 247-274), sur « les sources de la diversité » et « les racines du vernaculaire ».

2. Comparer dans l'éloigné, aires, villes, pays ou langues : ce que MPF a établi

Nous parlerons maintenant d'un corpus qui ne pose pas la comparabilité dès le dispositif d'enquête. La démarche à laquelle nous avons abouti s'est instituée peu à peu, en fonction des rétroactions par le terrain et d'ajustements en va-et-vient.

2.1. Les défis de comparabilité dans MPF

Le projet MPF[7] (*Multicultural Paris French*, voir Gadet et Guerin 2012) a commencé par une réflexion sur le recueil de données dans un corpus, au-delà de la reconduction de méthodes ayant (plus ou moins) fait leurs preuves ailleurs, sur d'autres terrains, avec d'autres protagonistes et d'autres objectifs. En mars 2015, MPF compte presque 700.000 mots transcrits/revus/anonymisés. Un tel recueil constitue une tâche suffisamment consommatrice de temps et d'énergie pour ne pas la contraindre à une méthodologie à laquelle il y a des raisons théoriques de ne plus adhérer.

Les objectifs du projet impliquaient de *comparer* : entre les enregistrements du corpus, mais aussi avec d'autres corpus, de la région parisienne ou d'ailleurs en France, en particulier concernant des jeunes en contexte urbain multiculturel, comme les données de Trimaille à Grenoble (parmi d'autres, Trimaille 2005 – comme beaucoup de sociolinguistes ancrés sur un terrain, Trimaille ne qualifie pas ses données de *corpus* – voir aussi Jamin *et al.* 2006, sur la comparaison entre régions de « formes supra-locales non standard », p. 338).

[7] ANR FR-09-FRBR_037-01, achevée depuis février 2014, dont le travail se poursuit sous d'autres formes.

Nous avons vite renoncé à sélectionner les « enquêtés »[8] en leur assignant de remplir les cellules d'une grille macro-sociologique, pour deux raisons au-delà des doutes théoriques sur l'intérêt d'un tel protocole. D'une part, les premiers objectifs d'exploitation (élargis par la suite) étaient syntaxiques et discursifs, et le socio-démographique manifeste probablement plus d'effet au niveau phonique, en relation plus immédiate avec le corps que ne l'est la grammaire (voir Armstrong 1998). D'autre part, l'objectif de recueil ordinaire (vernaculaire) contraint les circonstances de recueil, y compris dans les entretiens, surtout si on se laisse inspirer par une conception du style de « bon sens », comme effet d'auto-surveillance – ce qui conduit à conforter surtout les paramètres caractérisant un locuteur isolé, ou les traits superficiels de la situation.

Aussi avons-nous opéré une sélection sur une base communicative et interactionnelle, privilégiant l'échange entre interactants partageant une histoire conversationnelle. Même si l'idée de corpus implique des gestes peu propices au spontané (surtout dans les entretiens), étant donné le rôle actuel qu'ils revêtent en linguistique, nous voulions dépasser la dichotomie courante corpus (souvent assimilés à « entretien sociolinguistique ») *vs* recueil écologique.

Au-delà de la comparaison entre corpus de français, le grand nombre de travaux sur les parlers de jeunes « multiculturels » dans différentes grandes villes suggère des comparaisons, de ville à ville, de pays à pays, de langue à langue (de phénomène à phénomène ? de tendance à tendance ?). Les « parlers jeunes » constituent en effet une question vive, des points de vue historique, social, politique. Et pour comparer, il faudra établir quels objets comparer, parmi des méthodologies et des objectifs divers (voir Gadet et Hambye 2014 pour une bibliographie sur l'Europe)[9].

2.2. Comment MPF s'est confronté au défi de la comparabilité

Nous avons vu qu'une sélection des locuteurs reposant sur un échantillonnage socio-démographique n'assurait nullement la comparabilité affichée, dès lors illusoire. Avec pour hypothèse un poids de la qualité de l'interaction dans la diversification des façons de parler, nous avons privilégié la proximité (*vs* distance – voir Koch et Œsterreicher 2001), la densité de savoir partagé entre interactants, l'implicite permis par un réseau partagé. Le lien pouvait être antérieur au recueil (comme pour des assistants d'éducation s'entretenant avec des élèves),

[8] L'inadéquation de ce terme à la forme passive a souvent été soulignée. Mais les termes concurrents ont tous aussi des défauts plus ou moins rédhibitoires.

[9] Tous les termes permettant de qualifier les jeunes qui nous intéressent et leurs façons de parler manifestent le risque de pré-catégorisations. Il serait préférable de parvenir à en faire l'économie, et de se contenter d'une catégorie comme le « Vernaculaire Urbain Contemporain » de Rampton (entre autres 2011).

construit pour l'occasion (comme par cet enquêteur qui a fait de l'aide aux devoirs dans une association avant de sortir l'enregistreur) ou encore indirect, à travers un solide réseautage « d'ami d'un ami » (comme dans un entretien avec le jeune frère, jamais encore rencontré, d'un ami intime) – soit, selon la classification de Milroy and Llamas (2013), des réseaux de "first order" ou de "second order" (p. 411). Mais aucun enquêteur ne s'est entretenu avec des inconnus, ce qui nous permet de dire que nous avons fait des « entretiens de proximité ».

À chaque enquêteur, il a été demandé de mobiliser les ressources de ses réseaux afin de tendre vers une qualité sociolinguistiquement *audible*[10]. Un passage au crible des données après recueil repose sur un tamis intervenu à trois niveaux : 1) on enregistre des jeunes retenus sur leur façon de parler et pas seulement sur un profil socio-culturel ou démographique ; 2) on s'appuie sur une relation antérieure entre les protagonistes ou un solide truchement de mise en contact ; 3) un jugement après-coup (en équipe) évalue la qualité sociolinguistique audible, et les enregistrements peu convaincants sont écartés. Cette option sur la proximité a conduit à multiplier les enquêteurs, car il apparaît que les bonnes ressources d'un enquêteur sont en général épuisées après quelques enregistrements (ce qui est prévisible dans une perspective de réseaux). C'est pourquoi MPF est aujourd'hui le produit conjoint de 23 enquêteurs. Quelques réajustements ont été faits au fur et à mesure, quand on a pris conscience d'une question, comme l'impact, parmi une population de jeunes « beurs », de ce que plusieurs de nos enquêteurs étaient maghrébins (voir Gadet et Kaci à paraître). Ce qui nous a conduits à revisiter la réflexion sur la paire en interaction[11]. Toutefois, tout n'est pas amendable, et il faudra assumer les effets de quelques impensés (ou pensés trop tard).

Certains verraient sûrement cette diversité comme une inadmissible source d'hétérogène, comme si un enquêteur unique pouvait garantir une comparabilité sociolinguistique : tel n'est pas le cas, puisqu'un entretien est toujours une interaction – et *a fortiori* les événements écologiques. Car nous avons, quand cela s'est avéré réalisable (moins souvent qu'on ne l'avait espéré) redoublé les entretiens d'enregistrements écologiques recueillis par l'un des participants[12].

---

[10] Cette expression, pour le moment définie de façon intuitive, mérite d'être approfondie. Maintenant, quels sont les effets linguistiques d'une telle exigence, et pour quels phénomènes ? (sans doute pas pour tous). Ces questions sont à élaborer au cas par cas.

[11] Ce qui ne signifie pas qu'un enquêteur ne peut faire de « bons » entretiens qu'avec ses semblables, comme il est discuté dans Mayer (1995) qui revient sur les entretiens de Bourdieu et ses collaborateurs dans *la Misère du monde*. Ici non plus, il n'y a pas de principe universel : voir Becker (1998), qui se penche sur les conséquences de chacun des choix que fait le chercheur sur le terrain. Il serait toutefois illusoire de penser qu'on aurait fait l'économie de cette question en recourant à un enquêteur unique.

[12] Les auto-enregistrements avec des pairs, qui exigent un truchement, butent souvent sur le degré de compréhension de l'auxiliaire à l'égard de ce qui est attendu de lui. Nous avons ainsi dû en écarter un certain

Loin que les entretiens constituent un genre unique homogène, ils s'avèrent en continuum, certains se rapprochant d'une conversation ordinaire. Nous avons ainsi croisé une question vive en sociolinguistique, la mise en cause du lien posé dans « la première vague » d'études de la variation, entre identité et catégorisation étique (voir Eckert 2012, Brubacker and Cooper 2000, Kiesling 2013, qui tous mettent en cause la facilité éthique dans la définition de *identité*).

2.3. Le design des corpus

Les métadonnées décrivant la situation de recueil renseignent sur des paramètres objectifs (participants, cadre spatio-temporel, lien entre interactants, etc.), mais aussi sur des aspects plus subjectifs entrant en ligne de compte pour la qualité des données. La première série procure des informations permettant des statistiques (ce qui ne les rend pas *ipso facto* objectivables). Mais elle n'est pas à isoler d'une seconde série, qui échappe aux catégorisations et aux tentatives de formalisation, affine les premières informations et parfois ré-oriente l'interprétation au-delà de paramètres socio-démographiques. Ainsi de l'enregistrement d'une jeune femme que les indicateurs externes auraient posée comme parfait prototype de jeune de banlieue... sauf qu'elle a aimé l'école, et que ces deux facettes s'entrelacent dans sa façon de parler (entretien avec une intime). Aurait-elle rempli adéquatement une case « jeune de banlieue » ? Il est vrai que le fait de prendre appui sur des paramètres externes parie sur le recours au grand nombre pour amortir les idiosyncrasies – non sans prétendre au représentatif.

Si l'on tient compte de la qualité des échanges, comparer implique de ne pas se contenter de la façade (Kiesling 2013), comme le montrent les paramètres sous un paramètre. Un phénomène linguistique peut en effet répondre positivement à une corrélation, semblant ainsi sensible, par exemple à l'âge ou au sexe, alors qu'il peut ne s'agir que d'un épiphénomène d'un principe derrière l'âge ou le sexe. Les façons de parler des adolescents en constituent un bon exemple, car autant que leur âge biologique, leur sociabilité en réseaux serrés et cohésifs a des effets sociolinguistiques de contrôle de groupe (Kerswill 2010, Milroy and Llamas 2013, Gadet and Hambye 2014).

Certains jugeront sans doute une telle approche disparate, du fait de l'extrême diversité des interactants et des conditions de recueil, même si elles sont reconnues pour telles. Il y a justement là un point nodal pour la comparabilité, si du moins cette question n'est pas regardée comme réglée d'avance. Elle se pose peut-être encore plus vivement pour les vernaculaires, du fait qu'on a affaire à des normes qui se laissent difficilement saisir hors d'interactions au sein

---

nombre, jugés après-coup peu convaincants à l'écoute, dans un sens ou dans l'autre (soit surjoués dans l'ordinaire, soit reflétant des effets de pression normative).

d'un groupe – et c'est pourquoi les entretiens ont été faits, chaque fois que possible, avec des dyades ou des petits groupes : que cette pratique ait permis d'atteindre quelque chose de linguistiquement signifiant, c'est ce que cherchent à montrer ici-même Guerin et Moreno sur un phénomène très sensible à l'interaction, le discours rapporté.

Avec une telle conception d'évidence du social et de théorisations *ad hoc*[13], quel rôle est départi à « l'enquêté » ? Partenaire de plein exercice dans une véritable interaction ? Ou bien fournisseur passif de données, compte tenu de pré-catégorisations elles aussi prises pour évidentes, donc formulables par tout un chacun dont le non-expert du social qu'est le linguiste ? Dans une telle posture, on verra chez les jeunes des traits de « parler jeune », chez les Québécois des traits québécois, selon un processus d'*iconisation*, rançon des gommages (Irvine and Gal 2000).

Il n'y aurait alors pas à se soucier de choix individuel, ni d'agentivité des agents (entre agents libres et automates pré-déterminés pour la reproduction sociale), ni reconnaissance de perspectives émiques (Olivier de Sardan 1998).

3. Alors, quoi, comment et pourquoi comparer ? Remarques conclusives

Quiconque a fréquenté un terrain s'est confronté à une question ayant des incidences théoriques : quelle place pour l'inattendu/imprévisible (voir ici-même Gasquet-Cyrus) ? C'est là une limite incontournable des grands corpus : leur inévitable résistance, fort compréhensible dans leur logique propre, aux ajustements méthodologiques, même quand un chercheur est devenu conscient d'insuffisances du protocole – au nom de la comparabilité, qui impose de terminer un projet comme on l'a initié, quand ce n'est pas l'obligation institutionnelle de respecter un engagement. Or, qui dit terrain dit imprévisible.

Mais n'est-ce pas précisément pour cet imprévu que l'on va sur le terrain ? Pourquoi, sinon si aucune place ne devait être laissée pour ce prévisible imprévisible, sans lequel le terrain ne serait qu'un temps de vérification d'hypothèses et/ou de théories conçues préalablement ? Terrain et corpus semblent alors difficilement conciliables, le terrain ne pouvant recueillir, avant les démarches de généralisation, que du toujours spécifique, de l'indéfiniment diversifié (voir ici-même Hambye). Les (grands) corpus n'auraient ainsi aucune chance de montrer "why people behave linguistically as they have been found to do" (Cameron 1990, p. 81) – alors que

[13] Parmi ces propos épinglés dans des études sur les parlers jeunes : « exprimer une identité », « reproduire les normes d'un groupe », etc., types d'explications qui tournent en rond. On pourrait citer aussi la notion de *face*, maladroitement reprise du premier Goffman comportementaliste, qui ne fait que désigner un comportement sans l'expliquer. Sur l'identité, voir parmi d'autres Brubacker and Cooper (2000), Kiesling (2013).

cela pourrait justement être une raison pour en constituer. Ce qui soulève une question centrale pour la sociolinguistique, celle de la relation entre les façons de parler des individus et ce que l'on peut généraliser dans leurs regroupements, groupes, réseaux, communautés (éventuellement « de pratique »), sociétés, ensembles d'individus qui parlent une langue – question qui ne saurait être réglée par un avatar du « locuteur représentatif ».

Pour pouvoir comparer, il faut chercher à comprendre, à « expliquer »[14] les différences entre façons de parler, au-delà de corrélations trop souvent prises comme terme ultime d'un raisonnement (Cameron 1990). Mais qu'est-ce que comparer ? C'est sur ce que Remaud *et al.* (2012, p. 13) appellent « le geste comparatiste » que nous allons terminer. Ce geste ne va pas de soi, car il met en cause ce que le fait de comparer permet d'expliquer, non sans conséquences sur la façon de concevoir la sociolinguistique.

Nous avons vu que les vastes recueils de données ne prennent en général pas garde aux invisibilisations sur la comparabilité, qui ne fait pas objet de discussions. Les sociolinguistes devraient-ils s'abstenir de participer à ces recueils, du fait que les agendas de projets bien cadrés (les grands corpus) favoriseront toujours la masse de données aux dépens d'un « investissement en profondeur du terrain » (Blanchet 2007) ? Certes, ce n'est pas parce qu'un linguiste a besoin d'une masse de données qu'il doit être taraudé par des questionnements sociolinguistiques. Mais quel sera l'intérêt de cette masse, si elle est le produit d'une série de gommages ? Gommage des conditions d'émergence de discours situés, des porteurs de discours qui parlent en général pour dire quelque chose[15], de l'omniprésence de l'hétérogène et de l'instabilité que la sociolinguistique a justement contribué à montrer… Le fait est que ces catégories gommées n'intéressent pas la plupart des linguistes, qui ne veulent voir dans les corpus que de vastes « réservoirs de données » – mais avec quelles conséquences sur les données et quelles répercussions sur leurs analyses ? Un exemple parmi d'autres (bien peu nombreux) est l'article de Scheer (2013), qui soulève de bonnes objections à l'usage des corpus après être passé très vite sur leur constitution, en se contentant de rappeler après Saussure que le point de vue crée l'objet.

L'absence de prise en compte de ces questions pour les grands corpus laisse penser que les débats en sociolinguistique depuis les années 70 n'ont guère touché des démarches « de bon

[14] Les guillemets visent à instaurer une distance envers ce terme, omniprésent dans la littérature sociolinguistique qui le fait fonctionner à l'évidence. *Expliquer/explication* seront désormais utilisés sans guillemets, mais avec distance envers l'idée de corrélations.

[15] Blanche-Benveniste conclut son ouvrage de 2010 par un appel à poursuivre la collecte de données de français parlé : « rassembler des productions variées, dans des situations très diversifiées, traitées selon toutes les techniques modernes, mais sans que ces techniques empêchent de s'intéresser au contenu des textes » (p. 223). Elle n'ignorait sans aucun doute pas que ce propos était de l'ordre du vœu pieux.

sens », qui poursuivent sans état d'âme les pratiques antérieures – ce qui peut se comprendre, étant donné la force de l'évidence. Nous pensons cependant que les sociolinguistes peuvent jouer un rôle en montrant à quel point les observables constituent des données construites, donc un certain regard porté sur un certain terrain. Leur apport sur les corpus est de déconstruire chacun des gestes méthodologiques pratiqués lors des recueils (Vincent 2009 parle de « maîtriser tout le processus, de la conception de la recherche à l'offre d'accès aux corpus », p. 12). On n'attend pas d'eux qu'ils se limitent à faire des corpus, mais ils peuvent inciter à préférer des méthodes innovatrices à l'application répétitive de ce qui a été fait avant, même si on sait bien qu'il n'y a aucune chance qu'aucun corpus atteigne la perfection. Parce que les données des corpus, par essence même, sont construites, selon des procédures spécifiques, donc biaisées et partielles – ce qui ne veut pas dire nulles et non avenues, mais *situées*.

Une démarche sociolinguistique, aussi bien sur un terrain qu'en constituant des corpus (et pas forcément sans souci de terrain), de la première conception aux analyses, implique de la comparaison en arrière-plan (le « même » et le « différent »). Une *démythologisation* par le passage au crible des effets de chacun des gestes pratiqués sur le terrain devrait permettre à la sociolinguistique de conjoindre le terrain et la théorisation – et, pour ceux qui ne portent pas d'intérêt au sociolinguistique, d'améliorer la qualité des données sur lesquelles les linguistes élaborent désormais leur compréhension du langage et des langues.

Références

Armstrong N. (1998), « Perspectives sociolinguistiques sur la grammaire variable en français et en anglais », *Revue PArole* 3-4, p. 191-216.

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

Becker H. (1998), *Tricks of the Trade: How to Think about your Research while you're Doing It*, Chicago, University of Chicago Press.

Blanche-Benveniste C. (2010), *Le français. Usage de la langue parlée*, Peeters, Leuven & Paris.

Blanchet P. (2007), « Sur le statut épistémologique de la notion de 'corpus' dans un cadre ethno-sociolinguistique », dans Auzanneau M., dir., *La mise en œuvre des langues dans l'interaction*, Paris, L'Harmattan, p. 341-352.

Blondeau H., Sankoff G. et Charity A. (2002), « Parcours individuels dans deux changements linguistiques en cours en français montréalais », *Revue québécoise de linguistique*, Vol 31-1, p. 13-38.

Brubacker R. and Cooper F. (2000), "Beyond 'identity'", *Theory and Society* 29, p. 1-47.

Cappeau P. (2012), « L'intrigante évolution des paramètres (sociolinguistiques ?) utilisés dans les métadonnées », *Cahiers de linguistique* 38-1, p. 19-40.

Chambers J. (2003), *Sociolinguistic Theory: linguistic variation and its social significance*, Oxford, Blackwell (2nd éd.).

Chambers J., Trudgill P. and Schilling-Estes N. (2002), *The Handbook of Language Variation and Change*, Oxford, Blackwell Publishing.

Chambers J. and Schilling N. (2013), *The Handbook of Language Variation and Change*, Oxford, Blackwell Publishing, 2nd ed.

CIEL_F, Corpus International Ecologique de la Langue Française, http://ciel-f.org/

Coulmas F. (ed.) (1997), *The Handbook of Sociolinguistics*, Oxford, Blackwell.

Coupland N. (2007), *Style. Language Variation and Identity*, Cambridge, Cambridge University Press.

Gadet F. & Wachs S. (2015), « Comparer des données de corpus : évidence, illusion ou construction ? » Langage et Société 154, Maison des Sciences de L'homme Paris, 33-49.

DELIC (2004), « Présentation du *Corpus de référence du français parlé* », *Recherches sur le Français Parlé* 18, p. 11-42.

Durand J., Laks B. et Lyche C. (2009), « Le projet PFC (Phonologie du Français Contemporain) : une source de données primaire », dans Durand J., Laks B. et Lyche C., dirs., *Phonologie, variation et accents du français*, Paris, Hermès, p. 19-61.

Eckert P. (2012), "Three Waves of Variation Study: the Emergence of Meaning in the Study of Sociolinguistic Variation", *Annual Review of Anthropology* 41, p. 87-100.

ESLO http://eslo.in2p3.fr

Gadet F. (2011), "What can be learned about the grammar of French from corpora of French outside France", in Konopka M., Kubczak J., Mair C., Sticha F. and Wassner U., eds., *Grammatik und Corpora 2009*, Tübingen, Narr Verlag, p. 87-120.

Gadet F. (2013), *Banque de données sur les français hors de France*, Site de la DGLFLF.
http://www.dglflf.culture.gouv.fr/recherche/corpus_parole/BDD_Corpus_oraux_des_francais_hors_de_France

Gadet F. et Guerin E. (2012), « Des données pour étudier la variation. Petits gestes méthodologiques, gros effets », *Cahiers de linguistique* 38-1, p. 41-65.

Gadet F. and Hambye Ph. (2014), "Contact and Ethnicity in 'Youth Language' Description: In Search of Specificity", in Nicolaï R., ed., *Questioning Language Contact, Limits of Contact, Contact at its Limits*, Leiden-Boston, Brill.

Gadet F. et Kaci N. (à paraître), « Identification en première personne. Le discours d'un 'jeune de banlieue' en entretien », *Cahiers de Praxématique*.

Gadet F. *et al*. (2012), « CIEL_F : choix épistémologiques et réalisations empiriques d'un grand corpus de français parlé », *Revue Française de Linguistique Appliquée* XVII-1, p. 19-54.

Irvine J. and Gal S. (2000), "Language ideology and linguistic differentiation", in Kroskrity P., ed., *Regimes of Language*, Santa Fe: School of American Research, p. 35-84.

Jamin M., Trimaille C. et Gasquet-Cyrus M. (2006), « De la convergence dans la divergence : le cas des quartiers pluri-ethniques en France », *Journal of French Language Studies*, vol. 16-3, p. 335-356.

Kerswill P. (2010), "Youth Languages in Africa and in Europe: Linguistic Subversion or Emerging Vernaculars?", http://www.lancaster.ac.uk/fass/doc_library/linguistics/kerswill/Kerswill-African-Studies-19-10-10.pdf

Kiesling S. (2013), "Constructing Identity", in Chambers J. and Schilling N., eds., p. 448-467.

Koch P. et Œsterreicher W. (2001), « Langage oral et langage écrit », in Holtus G., Metzeltin M. et Schmitt C., eds., *Lexikon der romanistischen Linguistik*, Tome 1-2, , Tübingen, Max Niemeyer Verlag, p. 584-627.

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

Le Bianic T., Verdalle L. de et Vigour C. (2012), « S'inscrire dans une démarche comparative. Enjeux et controverses », *Terrains et Travaux* 21, p. 5-21.

Mayer N. (1995), « L'entretien selon Pierre Bourdieu. Analyse critique de *La Misère du monde* », *Revue française de sociologie*, 36-2, p. 355-370.

Meyerhoff M. (2006), *Introducing Sociolinguistics*, London & New York, Routledge.

Milroy L. and Llamas C. (2013), "Social Networks", in Chambers J. and Schilling N., eds., p. 409-427.

Moreau M.-L. (dir.) (1997), *Sociolinguistique. Concepts de base*, Sprimont, Mardaga.

Nicolaï R. et Ploog K. (2013), « Frontières. Question(s) de frontière(s) et frontière(s) en question(s) : des isoglosses à la 'mise en signification du monde' », dans Simonin J. et Wharton S., dirs., *Sociolinguistique du contact. Dictionnaire des termes et concepts*, Lyon, ENS Editions, p. 263-287.

Olivier de Sardan J.-P. (1998), « Émique », *L'Homme*, Persée 38-147, p. 151-66, http://www.persee.fr

PFC Phonologie du Français Contemporain, http://www.projet-pfc.net/?accueil:intro

Rampton B. (2011). "From 'Multi-ethnic adolescent heteroglossia' to 'Contemporary urban vernaculars'", *Language and Communication* 31: 276–294.

Remaud O., Schaub J.-F. et Thireau I. (2012), *Faire des sciences sociales. Comparer*, EHESS.

Scheer T. (2013), "The Corpus: A Tool among Others", *CORELA* [en ligne], HS15, mis en ligne le 19 février 2014, http://corela.revues.org/3006, Consulté le 25/03/2015.

Tagliamonte S. (2002), "Comparative sociolinguistics", in Chambers J. *et al*., p. 729-763.

Thibault P. et Vincent D. (1990), *Un corpus de français parlé. Montréal 84 : historique, méthodes et perspectives de recherche*, Québec, Département de langues et linguistique, Université Laval.

Trimaille C. (2005), « Spatialité vécue, dite et (inter)agie par des adolescents dans un quartier péricentral en mutation. Signalétiques et signalisations linguistiques et langagières des espaces de ville (configurations et enjeux sociolinguistiques) », *Revue de l'Université de Moncton* (Vol. 3-1), p. 61-96.

Vincent D. (2009), « Corpus, banques de données, collections d'exemples. Réflexions et expériences », *Cahiers de Linguistique* 33-2, p. 81-96.