Verbum XXX, 2008, 4

LES CORPUS ORAUX ET LA DIVERSITÉ DES PRODUCTIONS LANGAGIÈRES

Françoise GADET
Université de Paris Ouest Nanterre la Défense
& MoDyCo

RÉSUMÉ

Cet article revient sur la façon dont les grands corpus oraux ouvrent de nouvelles possibilités pour documenter des faits oraux de variation syntaxique en français. Il s'interroge sur les paramètres à la source de la diversification des productions langagières, comme les catégories socio-démographiques, dont l'usage est illustré dans l'échantillonnage des interviews. Il réfléchit ensuite aux caractéristiques des phénomènes variationnels syntaxiques, spécifiques parmi les faits variationnels.

ABSTRACT

This paper is concerned with the way corpora open new possibilities to document oral data for studying syntactic variation in French. Parameters involved in the diversification of discursive productions are then discussed, in particular the socio-demographic sampling as it is used in "sociolinguistic interviews". The second part of the paper deals with the differences between different syntactic phenomena: we try to explore how far syntactic data are specific among variationnal data.

Avant d'envisager de recueillir de nouveaux corpus oraux, toujours nécessaires, il est bon de prendre le temps d'évaluer les qualités et les limites de ceux qui sont actuellement disponibles, en particulier pour ce qui concerne les modalités de leur constitution. Il faut avant tout tenir compte des objectifs d'exploitation : sans négliger l'intérêt et la nécessité de corpus "tous objectifs", il est manifeste que le fait de poursuivre des objectifs phonologiques, syntaxiques, sociolinguistiques ou interactionnels[1] ne conduira

1 Toutes ces catégories mériteraient d'être précisées. Ainsi, pour des objectifs sociolinguistiques, selon qu'ils sont quantitatifs ou qualitatifs, ce ne seront pas les mêmes qualités qui seront requises des corpus.

pas aux mêmes exigences quant aux caractéristiques requises. Nous nous limiterons ici à considérer les premières étapes de recueils pour des objectifs syntaxiques ou discursifs avec des perspectives sociolinguistiques, avant même la phase de transcription (voir Cappeau & Gadet, à paraître, pour une réflexion sur ces phases ultérieures).

1. QUELQUES CONSIDÉRATIONS SOCIOLINGUISTIQUES

Il n'est pas inutile tout d'abord de revenir sur les typologies des données de corpus oraux qui sont couramment présentées, et de tenter de comprendre pourquoi certains types/genres discursifs s'avèrent si mal représentés.

1.1. Différents types de données

Selon Baude (2006, 26 sq.), les données documentées dans les corpus peuvent relever de deux types :

- des données "sollicitées" : il s'agit la plupart du temps d'interviews, souvent dites "interviews sociolinguistiques"[2] ;

- des données "de parole continue". Ce terme n'est lui-même pas univoque, un enregistrement "en continu" pouvant renvoyer aux productions d'un seul locuteur au cours d'un certain laps de temps, ou bien à ce que captent un enregistreur ou une caméra placés dans un lieu déterminé. La conséquence n'est pas la même quant à la saisie de l'événement discursif.

Il est donc souhaitable de spécifier l'expression "parole continue", en lui subsituant un autre terme : la dénomination "écologique" renvoie à des données dont l'existence sociale n'est pas liée à la seule quête du chercheur (dès lors provoquées ou organisées par le chercheur), parce qu'elles sont régulées par des fonctionnements sociaux qui auraient lieu même sans l'existence du chercheur et sa démarche de sollicitation, si "légère" soit-elle supposée. S'il peut se faire que la deuxième catégorie de Baude recoupe des données écologiques, tel n'est pas toujours le cas, l'idée de "continu" pouvant entrer en contradiction avec celle d'événement discursif.

Il faut d'ailleurs préciser que des recueils écologiques sont loin de constituer une démarche répandue[3], même en des temps qui font parler Vincent (2008) de "la multiplication à l'infini des possibilités d'accumulation de données", la seule limite étant d'ordre pratique : le passage obligé, lourd, exigeant et consommateur de temps, par la transcription.

1.2. Les données ordinaires

Parmi les corpus existants, on peut constater actuellement une proportion importante de données sollicitées, que ce soit parmi les données hexa-

2 Cette dénomination, pour être répandue, n'en est pas moins curieuse, venant caractériser des données en fonction de leur objectif d'exploitation.

3 Voir a contrario le projet CIEL_F. La qualité "écologique" n'est pas à confondre avec les exigences de "bonnes pratiques", évidemment toujours à l'ordre du jour (Baude, 2006).

gonales (Cappeau & Seijedo, 2005), ou dans les corpus de français hors de France (voir Cappeau & Gadet, 2007, pour un inventaire à date, et l'enquête en cours de Gadet, dont les résultats seront diffusés sur le site de la DGLFLF).

Les données qui s'avèrent les plus contraintes socialement, en tous cas au regard d'un point de vue d'observation extérieure, forcément étique (et non émique), donc les plus difficiles à recueillir, sont celles de zones discursives concernant les *vernaculaires*, sous la forme d'événements discursifs *ordinaires*. Le lien est évident avec le fait que, dans des conditions *écologiques*, ces données se manifestent dans des interactions ordinaires entre proches ou intimes. Les linguistes y font en général allusion à travers des termes qui méritent tous d'être précisés, comme *naturel*, *spontané*, *authentique*.

Labov (1976, 290) avait caractérisé le problème ici posé comme faisant l'objet du "paradoxe de l'observateur" ("[…] comment les gens parlent quand on ne les observe pas systématiquement, mais la seule façon d'y parvenir est de les observer systématiquement"), aux effets tellement nets qu'il lui est paru nécessaire de chercher des solutions pour le contourner. Nous allons montrer que cette question sociolinguistique débouche sur des questionnements qui concernent la grammaire : peut-on être sûr, à partir des seules interviews, d'accéder à tous les faits qu'on souhaite connaître et comprendre, en particulier ceux qui manifestent de l'hétérogénéité et de la variation ?

Cette question engage à réévaluer l'interview en tant que source de données. Ce n'est certes pas un genre discursif inconnu de la plupart des locuteurs, et elle fait bien partie de notre paysage communicatif actuel ; cependant, c'est plutôt au chapitre du contrôle social (Briggs, 2001). Ce qui a pour conséquence de fragiliser l'idée même d'un possible contournement ponctuel, puisqu'on ne sera pas davantage dans le registre du spontané.

1.3. L'interview sociolinguistique comme lieu de production de données

Etant donné le paradoxe de l'observateur, la méthodologie sociolinguistique classique a en effet développé la problématique de contourner, ou compenser, ce qui était regardé comme de simples "biais" de l'interview. Nous défendrons ici l'idée que cette problématique du contournement, fragile sur le plan sociologique car elle suppose possible de contourner la question sociale que désigne ce paradoxe (Gadet, 2002), offre l'intérêt d'amener le linguiste à réfléchir en termes de pondération entre les avantages et les inconvénients des divers gestes méthodologiques qu'il pratique. Elle nous conduit à poser deux questions, qui sont autant sociolinguistiques que grammaticales :

1) est-il socialement pensable de "désamorcer" ce que l'interview comporte de contrôle social, au-delà de bricolages ? N'y a-t-il pas naïveté sociologique à supposer un tel détournement possible ?

2) l'interview ne constituant évidemment pas la quasi-totalité (ni même la majorité) des productions discursives de nos sociétés, y a-t-il des consé-

quences à donner une telle importance à des données reflétant un événement discursif marginal ? Quels sont les effets de cette monotonie de genres discursifs ? Sont-ils négligeables, ou bien risquent-ils de se faire sentir dans la nature, la qualité, l'intérêt des données ? Y a-t-il un risque, avec cette limitation, de passer à côté de faits linguistiques, ou d'en mésinterpréter l'impact ou l'amplitude (fréquences, concordances, contraintes) ? Ces questions apparaissent particulièrement cruciales dans une perspective d'exploitation syntaxique des données.

Les conséquences de cette domination massive du genre *interview* peuvent ainsi se manifester à trois niveaux :

- effets du face-à-face (interaction à deux vs à plusieurs, et effet d'une éventuelle audience) ;

- effets de la relation d'interview, en particulier à cause de la dissymétrie entre questionneur et questionné ;

- effets de sous-représentation d'autres genres discursifs, qui auraient, peut-être, permis de voir autre chose.

1.4. Quelles sources pour une diversification des phénomènes linguistiques ?

Les grands corpus visent la plus grande diversité possible de formes de langue, pour des objectifs d'analyses aussi diversifiés que possible, et ils ont en général pensé la trouver en soumettant des locuteurs de profils socio-démographiques variés à un protocole qui les sollicite selon des modalités proches de l'interview[4]. Rechercher la plus large diversité de productions linguistiques, cette question oblige à interroger ce qui est à l'origine de la diversification des données : s'agit-il pour l'essentiel des locuteurs et de la diversité de leurs caractéristiques socio-démographiques ? de la palette des genres discursifs ? des situations (à supposer que l'on sache définir ce terme) ?

L'hypothèse la plus souvent avancée, au moins implicitement, est que la source de diversification majeure résiderait dans les différences entre locuteurs ; peut-être aussi est-ce la plus facile à appliquer. Cette conception frôle d'ailleurs à un tel point le stéréotype que, pour la plupart des linguistes, tel est le sens du terme *sociolinguistique*. Ce qui les conduit, plus ou moins implicitement, à des hypothèses de corrélation. Il semble alors inévitable de

4 Tel est le cas du *CRFP* (présenté dans le numéro 18 de *Recherches sur le Français Parlé*), autant que de *PFC* et de *CFPP* 2000. C'était aussi le cas du corpus de Montréal 1971 (voir Thibault & Vincent, 1990, et le regard inquisiteur qu'y jette aujourd'hui Vincent, 2008). Voir Gadet (2006) pour des interrogations sur les choix fondamentaux ainsi posés par la plupart de ces corpus. Toutefois, les positions d'interviewer ne sont pas dichotomiques, et certaines peuvent atténuer ces effets : voir entre autres Branca-Rosoff et al (à paraître en 2009, dont les interviews ont des objectifs plus crédibles que la simple volonté de "faire parler" ; ou le *CFPQ*, qui a privilégié la qualité des liens antérieurs avec les interviewés. Voir les réflexions de Cappeau & Gadet (à paraître).

suivre la pente de l'échantillonnage "représentatif", qui en ce cas sera à peu près immanquablement socio-démographique. Ces facteurs socio-démographiques sont d'ailleurs souvent qualifiés de "critères sociolinguistiques", alors même qu'ils n'ont à voir avec la langue que par la mise en corrélation. Il apparaît pourtant que ces seuls facteurs socio-démographiques, s'ils sont considérés de façon sèche, ne garantissent probablement pas toute la diversité linguistique souhaitée.

Pour que des critères socio-démographiques deviennent sociolinguistiques, il faut construire le problème sociolinguistique qui y correspond, comme on va le voir avec les catégories du sexe et de l'âge : aucun paramètre n'est en soi sociolinguistique, tant qu'il n'a pas été constitué comme tel.

1.5. Vous avez dit "représentatif" ?

Prenons d'abord le *sexe*. La plupart des études s'efforcent d'assurer une représentation égale d'hommes et de femmes. Or, c'est un critère auquel il est difficile d'attribuer du sens tant qu'on ne s'est pas demandé pourquoi le retenir, et avant d'avoir construit le problème sociolinguistique, qui n'est pas toujours le même selon les situations. La variabilité établie par différents travaux est considérable, de ce point de vue. Ainsi, l'étude de Coveney (2002) montre qu'il n'y a aucune incidence du sexe parmi le personnel d'une colonie de vacances, alors que celle d'Eckert (2000) affiche des différences majeures de comportements entre filles et garçons dans des bandes de jeunes organisées en réseaux. Il est clair toutefois qu'on ne peut éprouver l'incidence d'un critère que s'il a d'abord été retenu et testé. Et c'est pourquoi Milroy & Gordon (2003) conseillent de commencer par ces grandes catégories transversales à toutes les situations (âge, sexe, classe sociale, ethnicité…), dont il est ensuite facile de ne pas tenir compte si elles s'avèrent ne pas présenter d'intérêt.

Les autres catégories socio-démographiques fondamentales conduisent à des remarques similaires. L'âge ? Eckert (2000) a pu montrer que, chez les jeunes, le fonctionnement en réseau de pairs apparaissait plus important que l'âge. Quant à la localisation spatiale de l'habitat, il y a tout autant lieu d'interroger l'hypothèse derrière la diversification par un maillage géographique de villes, tel que mis en œuvre par exemple par le *CRFP*. On sait qu'il y a de nettes différentiations phonologiques, même si les accents régionaux ont été considérablement nivelés, comme l'a montré Armstrong (2001) : ils opposent essentiellement le nord et le sud de la France, mais pas spécialement l'intérieur de ces zones. C'est une autre question que de savoir dans quelle mesure les phénomènes syntaxiques et discursifs épousent le même type de schéma.

1.6. Il n'y a pas de représentativité identitaire

Ce goût pour le "représentatif", les linguistes devraient pourtant avoir appris à s'en méfier, au moins depuis que Gauchat (1905) a montré, dans une petite ville de Suisse romande, à quel point aucun locuteur ne parlait exactement comme les autres, même quand ils relevaient des mêmes caté-

gories socio-démographiques et professionnelles. De quoi en effet un locuteur pourrait-il être dit "représentatif", si personne n'est rapportable à une identité unique, un même individu pouvant par exemple être regardé à la fois comme femme, d'âge moyen, catholique, avocate, marseillaise, joueuse de bridge… ? Il n'y a que pour les démographes que les identités sont des catégories assignées à l'avance. Et les locuteurs ne sont pas porteurs d'une identité qu'ils revêtiraient à la demande afin de remplir une case de la grille du chercheur.

On peut à ce propos rappeler la notion de "collection" chez Sacks (1992) : une catégorie identitaire n'est susceptible d'être activée que de façon contextuelle et située, dans les interactions spécifiques menées avec des locuteurs spécifiques et dans des contextes spécifiques. Voir ainsi les critiques de la pré-catégorisation essentialiste chez Mendoza-Denton (2002), qui met en cause la notion d'identité et la faiblesse conceptuelle de sa mise en œuvre ordinaire.

C'est compte tenu des apories du mythe de la représentativité qu'est né le désir d'avoir plutôt affaire à des recueils de données écologiques[5], étant donné toute la diversité des genres discursifs.

2. DES CORPUS POUR UNE SYNTAXE SENSIBLE À LA VARIATION

Il ne sera pas question ici de tenir compte d'argumentations qui ont pourtant de l'importance pour le recueil de données : la différence, dans la perspective d'exploitation, entre visées illustratives et visées heuristiques (ce que Tognini-Bonelli, 2001, représente à travers les termes *corpus based* et *corpus driven*).

2.1. Objectifs phonologiques vs syntaxiques

Les faits phoniques et les faits syntaxiques n'imposent pas exactement le même type de défi à un concepteur de corpus. Mais la sociolinguistique, selon une conception ordinaire souvent maladroitement reflétée dans les corpus, continue d'avoir à gérer sur ce point l'héritage de ses origines dans la phonologie (et de ses attaches à une pensée de type structuraliste), ayant implicitement imposé une homologie des niveaux (ce qui vaut en phonologie vaudrait aussi en syntaxe).

Or, la disponibilité des faits linguistiques dans un corpus diverge bien, selon que les objectifs d'exploitation sont phonologiques ou syntaxiques.

2.1.1. La masse critique

Une première différence concerne la masse de données requises, selon au moins trois dimensions. C'est d'abord le nombre de phénomènes qui est

5 Cependant, on comprend bien que de telles conditions ne fassent pas l'affaire de tout type de chercheur. Ainsi, dans *PFC*, il est crucialement besoin d'assurer une comparabilité que l'on ne peut obtenir systématiquement que sur sollicitation (paires minimales et procédés de lecture, par exemple). C'est encore une raison pour insister sur le rôle majeur des objectifs de recherche.

en cause, plus limité en phonologie qu'en syntaxe (à condition de s'entendre sur ce que l'on appelle "phénomène", ce qui n'a rien à voir avec l'existence de contraintes, qui existent tout autant en phonologie qu'en syntaxe) - étant bien évident que les problèmes posés par l'intonation ne sont pas du même ordre.

La récurrence des phénomènes est en rapport avec leur nombre : toute transcription de français parlé exhibe en moyenne cinq ou six occasions de *e* muet par ligne (réalisé ou non) ; les occasions de chute ou de maintien d'une liquide post-consonantique sont déjà plus rares. En syntaxe, la palette des fréquences est très diversifiée : élevée pour la relation sujet-verbe ou pour les pronoms sujets, sans doute beaucoup moins pour les relatives, totalement liées au type de discours pour les négations, les interrogatives ou les concessives.

Par ailleurs, il n'y a à peu près pas de sensibilité au lexique en phonologie (sauf peut-être pour l'attaque en consonne ou voyelle), alors que c'est un trait majeur de la syntaxe, et cette propriété est sans doute encore accentuée à l'oral. Ainsi, Blanche-Benveniste (1997) écrit, à propos des constructions de *dont*, que ce pronom "a, par écrit, la moitié de ses emplois concentrés sur une dizaine de verbes, alors que, dans les conversations, le seul verbe *parler* accapare la moitié des emplois, le reste étant partagé par 8 autres verbes. Autant dire que, par oral, les emplois de *dont* sont presque entièrement stéréotypés, et qu'on pourrait en donner la liste plutôt que de les faire figurer dans une combinatoire libre de la grammaire" (1997, 74). Ce qui a des incidences pour des objectifs d'enseignement en français langue étrangère : *dont* ne serait pas prioritairement à enseigner en tant que structure susceptible de créativité.

2.1.2. Les possibilités d'évitement ou de contournement d'un phénomène ou d'une construction

Pour le plan phonique, il apparaît difficile de contourner toute occasion de liaison ou de *e* muet[6]. Il est au contraire possible d'éviter des subjonctifs ou des relatives, ou bien de modifier l'ordre des mots : en syntaxe, il y a toujours des solutions alternatives, même si elles engagent à de longues périphrases. On peut même dire que l'évitement constitue l'un des savoirs les plus sophistiqués sur le chemin de la maîtrise d'une langue (maternelle ou étrangère), qui ne vient qu'en couronnement d'un parcours, même si les apprenants en montrent très tôt des traces.

Accessoirement, on peut se demander dans quelle mesure les locuteurs sont conscients de ces corrélations : c'est probablement davantage le cas aux niveaux syntaxique et discursif qu'au niveau phonique.

---

6 C'est ce que Georges Perec est parvenu à faire dans son roman *La disparition*, mais il s'agit d'écrit, et il ne faut pas négliger la différence avec l'oral quant à la temporalité : le scripteur dispose de temps par rapport à un parleur, écrire étant plus lent que parler, et permettant la réflexion et la correction avant que ne soit livré le produit fini. Sur les implications linguistiques de ces données matérielles, voir Chafe (1985).

2.2. La sensibilité aux genres discursifs

Elle est limitée en phonologie, où il y a surtout des styles (pour simplifier, "formel" vs "informel", ce qui permet par exemple d'anticiper que l'on rencontrera davantage de e muets ou de liaisons dans un style formel). Il n'y aurait aucun sens à supposer que tel genre discursif pourrait favoriser ou empêcher la présence de e muets (sauf en verlan !). En syntaxe au contraire, se manifestent des incidences à la fois des styles et des genres. Nous allons maintenant en étudier quelques exemples.

2.2.1. L'exemple des participes présents

Certaines incidences des genres sont de l'ordre de l'évidence : il n'y a pas lieu de s'étonner, par exemple, de ne pas trouver d'interrogatives dans les réponses à une interview.

De façon moins nécessairement attendue, Bilger & Cappeau (2004) ont montré par exemple qu'un genre comme la "visite commentée" favorisait la présence de participes présents, en particulier en position initiale de séquence comme dans l'exemple (2) :

(1) *toutes les couches archéologiques correspondant à la date…* (p. 16)

(2) *pensant que le roi n'aurait pas été d'accord avec ce mariage…* (p. 16)

2.2.2. L'exemple de *nous* en position sujet

On sait à quel point *nous* en position sujet est rare en français contemporain parlé ordinaire[7], où il est généralement remplacé par *on*. Bilger & Cappeau (2004) ont montré que l'une des rares circonstances où l'on en rencontre beaucoup était dans un genre qui n'est pas couramment envisagé par les typologies : le "récit de voyage collectif" (à distinguer du récit de voyage tout court) :

(3) *Nous sommes arrivés à Calais après avoir roulé…* (p. 19)

On peut d'ailleurs rappeler à cette occasion que les regroupements discursifs auxquels parvient Biber (1988), sur la base de faisceaux de traits dont une bonne partie sont syntaxiques, ne recoupent pas fondamentalement les genres retenus par la tradition.

2.2.3. L'exemple des sujets nominaux

Blanche-Benveniste (2008, 24) qualifie le traitement des sujets nominaux de "pierre de touche pour juger du degré de formalité des discours". Son article de 1994 a montré que des syntagmes nominaux pleins en position

7 C'est justement pourquoi il est immanquablement perçu comme marque d'un discours surveillé. Ainsi, Lambert (2005) montre que des adolescents de "quartiers difficiles", placés en situation d'avoir à faire dans la rue des interviews de passants inconnus, exhibent des ressources imprévues, comme des liaisons facultatives et des *nous* en position sujet ; et en tous cas une parfaite politesse.

sujet (vs clitiques, en nombre bien plus élevé à l'oral, mais aussi vs dislocations) ne se rencontraient couramment que dans certains genres oraux spécifiques, comme les "explications techniques" :

(4) *Le procédé de fabrication est le même*

(5) *Les frais de dossier à l'ouverture sont faibles* (in Blanche-Benveniste, 2008, 22)

2.2.4. L'exemple des subordonnées

Pour étudier certains phénomènes comme les subordonnées, déjà faut-il que les séquences soient suffisamment longues pour en comporter. Mais les interactions ordinaires attestent que la parole circule très vite entre les interactants, laissant apparaître des unités plutôt brèves[8], souvent dépourvues d'élaboration syntaxique quoique pas forcément de complexité syntaxique, dans lesquelles il n'y a que quelques rares occasions de séquences longues. Les récits, genre où il est socialement convenu de laisser le locuteur en place garder longuement la parole, ne sont pas de bons candidats, parce que les structures y sont assez monotones et ne comportent pas forcément beaucoup de subordonnées. La seule véritable opportunité de longueur, ce sont les explications et les argumentations, voire les réfutations, qui provoquent des énoncés un peu longs et davantage d'imbrications syntaxiques.

2.3. Peut-on provoquer la production de phénomènes grammaticaux ?

Pour disposer d'un nombre élevé d'occurrences de tel ou tel phénomène, ne serait-ce qu'afin de déterminer les contraintes qui pèsent dessus, il n'y a guère que deux solutions : la mitraillette et la mise au point. Dans le premier cas, on compte sur la masse : si le corpus est suffisamment vaste et diversifié, il finira bien par émerger un certain nombre d'exemples pertinents, et c'est la tactique "arroser la cible". Dans le second cas, il s'agit de concevoir précisément des moyens de "provoquer" (*to elicite*) spécifiquement la production de telle ou telle forme grammaticale. Stratégie qui est d'autant moins à préconiser que toutes les catégories ne s'y prêtent pas.

De façon générale, grammairiens et linguistes ont fait des hypothèses sur la relation entre certaines tâches et l'émergence de certaines catégories grammaticales, comme on l'a vu avec les explications ou les argumentations. Il faudrait d'ailleurs aussi tenir compte, dans la diversité des situations, de ce que les protagonistes de l'événement discursif partagent ou non l'environnement d'un objet, d'une machine, d'un écran (avec quels effets syntaxiques, au-delà des déictiques ?). Mais de tels inventaires des ressources syntaxiques en liaison avec les situations discursives ou des genres, que Blanche-Benveniste (2008, 29) appelle de ses vœux, sont loin d'être disponibles en nombre suffisant.

---

8 Ainsi, dans la base de données CLAPI (*Corpus de LAngue Parlée en Interaction*), les visées interactionnelles peuvent ne pas convenir aux besoins d'un syntacticien, du moins pour certains phénomènes ou structures (en particulier dès qu'il est besoin d'énoncés longs). Cette remarque conduit à un questionnement en arrière-plan, concernant la relation entre corpus recueillis préférentiellement et objectifs d'exploitation visés. Voir entre autres Cappeau & Gadet (à paraître).

2.4. Granularité plus fine

Pour des sous-catégories grammaticales, on peut penser à l'exemple des futurs, simple ou périphrastique. Il semble à première vue facile, lors d'une interview sociolinguistique, de provoquer la présence de temps du futur, par exemple en faisant parler les interviewés de projets d'avenir ; on peut alors espérer recueillir suffisamment de matière pour des hypothèses sur les contraintes en relation avec la présence de futurs simples ou périphrastiques. Mais comment être sûr que le genre interview n'aura pas favorisé justement l'apparition d'une forme ou de l'autre ? Car comment documenter le phénomène avant d'en avoir fait l'analyse ? Les hypothèses sur les deux futurs, héritées de la grammaire traditionnelle, s'appuient en général sur des différences de sens, comme dans les annonces préenregistrées successives (6) et (6'), que je dois au grand usager des trains qu'est Paul Cappeau, - qui pourraient d'ailleurs bien refléter surtout une conformité au modèle scolaire, plus qu'un usage courant. Mais certains corpus permettent d'ouvrir sur d'autres hypothèses : ainsi, dans le corpus Deshaies (Québec, voir Deshaies & Laforge, 1981), on rencontre des exemples comme celui ici reproduit en (7), qui incite à lier le futur périphrastique à l'affirmatif, et le futur synthétique à la négation[9] :

(6) *le train numéro XX entrera en gare voie Y*

(6') *le train numéro XX va entrer en gare voie Y*

(7) *un jour je me dis je vais le faire / le lendemain, je me dis non je ferai pas ça*

Outre les temps, dont il est facile pour des raisons sémantiques de déclencher l'emploi (ainsi, des exemples de formes de temps du passé s'obtiennent à travers des récits, genre discursif facile à solliciter de façon naturelle), quelles autres catégories grammaticales pourrait-on ainsi provoquer, en tant que catégorie large, autant que dans le raffinement sémantique ? On peut prendre l'exemple des prépositions : on peut faire produire des prépositions locatives par exemple avec des descriptions d'itinéraires, ou bien des prépositions temporelles en provoquant des récits. Mais comment assurer la présence de prépositions vides, comme *à*, *de*, ou *pour ?* Et, au-delà, comment assurer le plus grand nombre de pronoms, de relatives, d'infinitifs… ? Ou encore une grande diversité dans l'ordre des mots ?

Il n'y a que quelques cas, finalement assez peu nombreux, pour lesquels on peut espérer qu'une documentation suffisante émergera de ce qui aura été provoqué.

2.5. Corpus et co-texte discursif large

Parmi les nombreuses raisons pour lesquelles il apparaît souhaitable de disposer de corpus diversifiés et de longues plages de texte plus que

9 On ne dispose pas pour le moment de suffisamment de corpus diversifiés pour argumenter si s'agit là d'un trait spécifiquement québécois, ou de tendances qui n'ont pas encore été perçues en français de France.

d'exemples isolés, nous n'en illustrerons qu'une, qui ouvre sur l'analyse de discours et laisse voir la sensibilité aux genres (comme descriptions de consignes, recettes ou explications techniques vs récits, par exemple). Blanche-Benveniste (1986) évoque la prise en compte d'"espaces de textes", dont il est impossible de prévoir la taille utile pour disposer d'exemples de "multi-formulations". Elle s'interroge sur le rôle de reformulations d'un même verbe comportant un même matériel lexical, sous les formes de l'actif, du passif et de formes verbales en *se,* comme dans les exemples suivants[10] :

(8) *la pellicule tu l'enroules bon une fois que c'est enroulé tu l'enfermes dans la boîte*

(9) *alors ça se dissout à l'alcool alors il y a plusieurs moyens il y en a qui dissolvent cette résine à l'alcool*

Ces reformulations viennent préciser, par le moyen de touches successives, la valeur d'une formulation, en particulier sur les constructions en *se*, qui peuvent être ambiguës, entre réflexif, réciproque, moyen ou passif. Les différentes formulations ainsi déployées les unes à côté des autres sur un axe syntagmatique peuvent s'entendre comme des indices paradigmatiques, permettant au chercheur de classer les constructions verbales, mais aussi ayant permis au locuteur de mettre au point sa recherche de formulation précise.

CONCLUSION

Les corpus multi-objectifs ne s'avèrent, certes, jamais tout à fait satisfaisants, quel que soit le phénomène syntaxique spécifique sur lequel on travaille. Pourtant, il serait déraisonnable de renoncer à en constituer, et même à en constituer constamment de nouveaux avec des bases de recueil davantage concertées et systématiques. Dans cette optique, il faut rappeler à quel point tout geste méthodologique, depuis les premières options pour le protocole de collecte jusqu'à l'analyse (sans négliger les modalités de collecte et de transcription), revêt une importance décisive (Milroy & Gordon, 2003). Si, dans la collecte des données comme dans les premières étapes de leur exploitation, un geste n'a pas fait l'objet de décisions mûrement pesées adaptées à la fois à la situation de recueil et aux objectifs visés, il risque fort de déclencher des effets indésirables parce que demeurés implicites.

Dans la "nouvelle façon de faire de la linguistique" que permettent aujourd'hui les grands corpus, il faudrait, idéalement, ne laisser pour les options fondamentales que le moins de place possible au hasard. C'est une condition indispensable pour avoir des chances de faire une description fiable du français parlé jusque dans des phénomènes "émergents" (Hopper, 1987), donnant ainsi toute son ampleur heuristique à une perspective inspirée de l'approche *corpus-driven* (Tognini-Bonelli, 2001).

---

10 Dans ces exemples, les deux formes se succèdent très rapidement, mais rien n'empêche qu'elles figurent à distance.

BIBLIOGRAPHIE

ARMSTRONG N. (2001), *Social and Stylistic Variation in Spoken French. A comparative approach*, Amsterdam, John Benjamins Publ. Company.

BAUDE O. (dir.) (2006), *Corpus oraux. Guide des bonnes pratiques*, Presses Universitaires d'Orléans/CNRS Editions.

BIBER D. (1988), *Variation across spoken and written language*, Cambridge, Cambridge University Press.

BILGER M. & P. CAPPEAU (2004), "L'oral ou la multiplication des styles", *Langage & Société*, 109, 13-30.

BLANCHE-BENVENISTE C. (1986), "La notion de contexte dans l'analyse syntaxique des productions orales : exemples des verbes actifs et passifs", *Recherches sur le Français Parlé*, 8, 39-57.

BLANCHE-BENVENISTE C. (1994), "Quelques caractéristiques grammaticales des 'sujets' employés dans le français parlé des conversations", in *Actes du Colloque Subjecthood and subjectivity*, Paris/Londres, Ophrys & Institut français du Royaume-Uni, 77-107.

BLANCHE-BENVENISTE C. (1997), *Approches de la langue parlée en français*, Paris, Ophrys.

BLANCHE-BENVENISTE C. (2008), "Le français parlé au 21e siècle. Réflexions sur les méthodes de description : système et variation", in Abecassis M., Assoyo L. & Vialleton E. (dir.), *Le français parlé au XXIe siècle. Norme et variations géographiques et sociales*, Paris, L'Harmattan, 17-39.

BRANCA-ROSOFF S., S. FLEURY, F. LEFEUVRE & M. PIRES (à paraître en 2009), *Constitution et exploitation d'un corpus de français parlé parisien*, http://ed268.univ-paris3.fr/CFPP2000/

BRIGGS C. (2001), "Interviewing, Power/knowledge, and Social Inequality", in J. Gubrium & J. Holstein (eds), *Handbook of Interview Research: Context and Method*, Thousand Oaks/London, Sage Publ.

CAPPEAU P. & F. GADET (2007), "Où en sont les corpus de français parlé ?", *RFLA* XII-1, 129-33.

CAPPEAU P. & F. GADET (à paraître), "Transcrire, ponctuer, découper l'oral : bien plus que de simples choix techniques", *Cahiers de linguistique*.

CAPPEAU P. & M. SEIJEDO (2005), *Inventaire des corpus oraux en langue française*, www.dglflf.culture.gouv.fr

CHAFE W. (1985), "Linguistic differences produced by differences between speaking and writing", in D. Olson, N. Torrance & A. Hildyard (eds), *Literacy, Language and Learning*, Cambridge University Press, 105-23.

CIEL_F : Corpus International Ecologique de la Langue Française, www.ciel-f.net

CLAPI : Corpus de Langue Parlée en Interaction, http://clapi.univ-lyon2.fr

CFPP 2000 : Corpus de Français Parlé Parisien des années 2000, http://ed268.univ-paris3.fr/CFPP2000/

CFPQ : Corpus de Français Parlé Québécois, http://pages.usherbrooke.ca/cfpq/index.php

COVENEY A. (2002) *Variability in Spoken French. A Sociolinguistic Study of Interrogation and Negation*, Exeter, Elm Bank Publications [2e édition avec postface].

CRFP : Corpus de référence du français parlé, voir *Recherches sur le Français Parlé*.

DESHAIES D. & E. LAFORGE (1981), "Le futur simple et le futur proche dans le français parlé de la ville de Québec", *Langues et linguistique*, 7, 23-37.

ECKERT P. (2000), *Linguistic Variation as Social Practice*, Oxford, Blackwell.

GADET F. (2002), "Derrière les problèmes méthodologiques du recueil des données", http://www.revue-texto.net/Inedits/Gadet_Principes.html

GADET F. (2006), Compte-rendu de *Recherche sur le français parlé* n° 18, *Bulletin de la Société de Linguistique de Paris*, tome CI, fasc 2.

GAUCHAT L. (1905) "L'unité phonétique dans le patois d'une commune", in *Aus Romanischen Sprachen und Literaturen: Festschrift Heinrich Mort*, Halle, Max Niemeyer, 175-232.

HOPPER P. (1987), "Emergent grammar", *Berkeley Linguistic Society* 13, 139-57.

LABOV W. (1972), *Sociolinguistic Patterns*, tr. fr. *Sociolinguistique*, 1976, Paris, Editions de Minuit.

LAMBERT P. (2005), *Les répertoires plurilectaux de jeunes filles d'un lycée professionnel : une approche sociolinguistique ethnographique*, thèse de doctorat, Université de Grenoble.

MENDOZA-DENTON N. (2002), "Language and Identity", in Chambers & al., *The Handbook of Language Variation and Change*, Oxford, Blackwell Publishing, 475-99.

MILROY L. & M. GORDON (2003), *Sociolinguistics. Method and interpretation*, Malden/Oxford, Blackwell Publishing.

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

Recherches sur le Français Parlé (2004), "Autour du corpus de référence du français parlé", n° 18, 11-42.

SACKS H. (1992), *Lectures on conversation*, London, Blackwell, 2 vol.

THIBAULT P. & D. VINCENT (1990), *Un corpus de français parlé. Montréal 84 : historique, méthodes et perspectives de recherche*, Recherches sociolinguistiques 1, Bibliothèque nationale du Québec.

TOGNINI-BONELLI E. (2001), *Corpus linguistics at work*, Amsterdam, John Benjamins.

VINCENT D. (2008), "Corpus, banques de données, collections d'exemples. Réflexions et expériences", *Cahiers de Linguistique*, 33/2, 81-96.