Réflexions sur un obscur objet de désir: le corpus

Introduction

Comme le signalent Cappeau et Gadet (2007), il semble y avoir «une double évidence» partagée aujourd'hui par la grande majorité des linguistes:

– d'une part: «il faut travailler sur l'oral»,
– d'autre part, il faut «le faire sur des corpus, les plus grands possibles».

L'objectif de cet article sera donc d'interroger ces deux évidences, et pour ce faire, je m'appuierai, entre autres, sur l'expérience acquise au cours de la constitution et de l'exploitation de deux types de corpus, celui que l'on appelle: le corpus du Groupe Aixois de Recherche en Syntaxe (GARS) ou encore CorpAix et celui du Corpus de Référence du Français Parlé (CRFP), qui a été piloté par l'équipe DELIC (DEscription Linguistique et Informatique sur Corpus, Université de Provence).

Je serai également amenée à rappeler les différentes conceptions que l'on peut avoir de cet objet, ce qui me permettra de revenir sur les caractéristiques, les avantages et les limites des corpus dits «textuels» et des corpus dits «de référence».

Je ne chercherai pas à justifier l'intérêt de travailler sur l'oral pour décrire le fonctionnement d'une langue. En revanche, je montrerai que l'opposition oral/écrit se révèle de moins en moins pertinente à la lumière des données attestées, si ces dernières sont suffisamment diversifiées.

Enfin, pour conclure, j'essaierai de répondre à la question que certains d'entre nous commencent à se poser, quant à l'apport réel de l'utilisation des corpus dans les sciences du langage.

1. Bref bilan de la politique des corpus en France

Aux cours de ces 10 dernières années, les linguistes français mais aussi les institutions ont pris conscience de la nécessité et de l'urgence qu'il y avait de constituer de grands corpus de français parlé. Tout le monde reconnaissait qu'il fallait combler le retard que le français de France avait pris dans ce domaine, par rapport aux grandes langues européennes et intercontinentales, par rapport aussi aux français de la francophonie (cf., entre autres, la banque de données de VALIBEL (Variétés Linguistiques du français en Belgique, Université catholique de Louvain-la-Neuve).

Ce constat a donc permis que d'importants projets fondés sur corpus aient été réalisés avec le soutien des institutions, entre autres, le corpus PFC (*Phonologie du Français Contemporain,* coordonné par Durand, Laks et Lyche, voir Durand et Lyche, 2003) ou encore le CRFP, qui s'est finalisé en 2004.

Pour favoriser le recueil des corpus, la Délégation Générale à Langue Française et aux Langues de France a par ailleurs financé et piloté, très récemment, deux types de travaux:

- d'une part, un inventaire des corpus de français parlé, existant en France et à l'étranger, afin d'identifier les manques les plus évidents dans le domaine des données orales, dans le but aussi d'aider les futurs projets à mieux cerner les forces disponibles et les besoins. Cet inventaire élaboré par Cappeau et Seijedo est disponible depuis 2005 (sur le site de la DGLFLF) et a été réactualisé cette année. Il a révélé, non pas un manque de corpus, mais un réel éparpillement, en ce qui concerne, entre autres, la nature et le format des données, ainsi que leur accessibilité.
- d'autre part, et sans doute pour remédier à cet éparpillement, la DGLFLF a publié en 2006 sous la direction d'Olivier Baude: «un guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux». Cet ouvrage a pour objectifs d'éclairer la démarche des chercheurs, de lister les problèmes et les solutions juridiques concernant le droit de la «parole», mais aussi et surtout de favoriser l'émergence de pratiques communes.

Enfin, depuis septembre 2006, la volonté de constituer «une grande banque de données orales» a été clairement exprimée par les autorités, puisque, sous l'égide cette fois du CNRS, a été créé et mis en place le Centre de Ressources des Données Orales (le CRDO).

Ce centre a pour but de regrouper les corpus constitués par des équipes différentes, ce qui permettrait de disposer rapidement d'un gros corpus accessible par le Net. Ce serait, en quelque sorte, l'équivalent «oral» du grand corpus écrit «Frantext».

Au premier abord, cette initiative peut paraître intéressante. Il est vrai qu'elle offre une solution rapide pour remédier à l'état actuel des besoins. Cependant, elle pose un certain nombre de problèmes dans la mesure où vont être regroupés des objets d'études fort disparates, en ce qui concerne la conception d'origine, les objectifs, mais aussi et surtout la façon dont ils seront édités.

De fait, ce projet semble oublier un peu vite que la constitution d'un corpus de langue orale engage et suppose une réflexion à la fois théorique et méthodologique sur les données.

C'est d'ailleurs ce poids de la théorie sur la pratique qui explique en grande partie pourquoi il y a une telle variété, en ce qui concerne, entre autres, les conventions de transcription utilisées, autrement dit, en ce qui concerne l'édition même de ces corpus. Par exemple, certains vont utiliser l'orthographe standard, d'autres une orthographe aménagée; certains vont utiliser les signes de ponctuation, d'autres pas, et la liste des divergences en la matière est grande.

Alors une question se pose: Est-ce que le chercheur pourra exploiter l'ensemble de ce grand corpus, en ignorant la façon dont chaque sous-partie a été conçue et sans qu'il y ait eu au moins un travail préalable pour harmoniser l'aspect matériel de ces données?

En fait, et on peut le regretter, il n'est jamais facile d'exploiter les corpus édités par d'autres équipes que la sienne – ou il y a trop d'informations, ou il n'y en a pas assez, ou il n'y a pas celles qui intéressent – bref, il semble bien que, même si la grande majorité des chercheurs le souhaitent, la mise en commun des données ne soit pas aussi simple que cela.

D'autre part, ce projet semble signifier que, pour pouvoir «faire vite, faire gros et pas trop cher», la tradition française qui a souvent favorisé le «corpus textuel» risque de se perpétuer, au détriment, bien entendu, du «corpus de référence».

2. Corpus textuel/corpus de référence, quelles différences?

De fait, le terme «corpus » renvoie à des réalités bien différentes qui ne cessent d'ailleurs d'évoluer (cf. Meyer, 2002; Rastier, 2005). En dehors du cas particulier où le terme «corpus» est utilisé pour désigner une simple collection de données, sous la forme d'un exemplier construit, ce terme renvoie de manière plus générale à l'idée d'un regroupement de textes permettant de mener à bien des recherches spécifiques.

Il existe cependant aussi bien en ce qui concerne la notion de «regroupement» que la notion de «texte» des pratiques ou des traditions divergentes. Aussi, oppose-t-on souvent «corpus textuel» et «corpus de référence».

Le corpus dit «textuel» est un corpus «ouvert», constitué de textes entiers. C'est par exemple pour le domaine du français écrit, la base Frantext[1] qui regroupe, entre autres, certains textes littéraires dans leur intégralité, du moyen français à nos jours, et qui s'enrichit régulièrement.

Cette conception du «corpus» constitué de textes entiers non organisés est celle qui a prévalu en France et qui, comme je l'ai dit, risque de se perpétuer avec le projet CRDO. A noter d'ailleurs que, dans ce cas, aujourd'hui, on parle plutôt «d'archive» que de «corpus» en tant que tel (cf. Rastier, 2005).

En revanche, dans les pays anglo-saxons, c'est la constitution de corpus dit «de référence» qui a été favorisée. Ce type de corpus se caractérise avant tout par le souci de présenter de manière équilibrée, sous la forme d'échantillons,[2] le plus grand nombre d'usages possible. Mais comme le signale Habert (2000:13) dans les deux cas, il s'agit plutôt de «base textuelle», de «réservoirs à corpus», dans la mesure où c'est seulement:

[1] Cf. pour une présentation détaillée de cette base de données, voir Martin (1996).

> l'opération de choix raisonné parmi les composants disponibles (dans la base en question) qui crée un corpus et qui permet de mener à bien une recherche particulière.

Habert propose d'ailleurs de modifier quelque peu la définition du terme «corpus» donnée par Sinclair (1996:4):[3]

> Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques **et extra-linguistiques** explicites pour servir d'échantillon ~~du langage~~ **d'emplois déterminés d'une langue**.

Ces deux types de regroupement de textes, de «corpus» présentent chacun des avantages et des inconvénients; Sinclair (1991) note qu'avec l'inclusion de documents complets, deux inconvénients sont évités:

- La sur-représentation de certaines parties des textes comme, par exemple, la partie initiale, «l'introduction», qui peut modifier les résultats concernant la fréquence de tel ou tel phénomène linguistique, comme l'a signalé Biber (1995).

- Les doutes que l'on peut toujours avoir sur la validité même des techniques d'échantillonnage.[4]

Mais Sinclair reconnaît aussi qu'avec le corpus «textuel», on est rapidement confronté au problème de la représentativité. Tant il est vrai qu'avec ce type de corpus, la couverture initiale est rarement aussi complète que celle que l'on peut obtenir à partir d'une collection planifiée d'échantillons, c'est-à-dire, à partir d'un corpus de «référence».[5] Voici la définition qu'en donne l'auteur (1996), cité par Habert (2000):

> Un corpus de référence est conçu pour fournir une information en profondeur d'une langue. Il vise à être suffisamment grand pour représenter toutes les variétés pertinentes de cette langue et de son vocabulaire, de manière à pouvoir servir de base à des grammaires, des dictionnaires et à d'autres usuels fiables.

Ce type de corpus est donc conçu pour que soient représentés de manière «équilibrée» les différents usages attestés dans diverses activités de langage. Et c'est cette exigence liée à la «diversité» qui rend difficile la constitution d'un tel outil.

---

[2] La taille des extraits varie selon les corpus et peut aller de 2000 mots (cf. les corpus de Brown et LOB (Lancaster, Oslo, Bergen) constitués de 500 extraits de 2000 mots chacun) à des extraits de 40 000 mots (cf. le corpus Longman/Lancaster English Language Corpus).

[3] Les termes en caractères «gras» ainsi que la séquence raturée correspondent aux changements proposés par. Habert (2000).

[4] De fait, comme le rappellent Francard, Geron et Wilmet (2002), il existe un certain nombre de problèmes concernant cette notion «d'échantillon»: difficulté pour savoir si les échantillons retenus reflètent bien de manière proportionnelle les différentes productions d'une population donnée; difficulté de trouver un consensus sur la dimension même de l'échantillon à partir de laquelle des régularités significatives commenceraient à apparaître.

[5] Cf. à ce sujet la critique du corpus Frantext faite par Rastier (2000).

L'idéal d'exhaustivité ne se trouve plus seulement dans la quantité des données recueillies, dans la taille même du «corpus», mais bien dans la représentation de tous les usages de la langue écrite et orale.

Les données du corpus de référence doivent donc, toujours dans l'idéal, être représentatives de la langue commune et générale, prise comme un tout.

En ce sens, aucun «genre littéraire» particulier ne doit dominer, ni aucun style ou forme de parler. Au contraire, les textes ou les extraits de textes doivent être hétérogènes et renvoyer au plus grand nombre d'usages possible.

De prime abord, ces objectifs peuvent sembler faciles à atteindre, du moins en grande partie. Cela pose cependant un certain nombre de questions qu'il est intéressant de rappeler.

En fait, trois problèmes essentiels se posent à l'heure de constituer ce type de corpus:

– Les fonctions du corpus
– Les supports sélectionnés
– Les genres de textes représentés

En ce qui concerne les fonctions du corpus, il va de soi que plus les utilisateurs envisagés sont variés, plus le corpus doit être diversifié, afin qu'il puisse répondre à toutes les attentes et satisfaire toutes les demandes. Les supports peuvent être également multiples: livres, périodiques, documents publiés ou non, lettres, dépliants, transcriptions de l'oral, etc.

Enfin, en ce qui concerne les «genres» de textes, la question semble plus complexe dans la mesure où la classification, qui se fait souvent *a priori*, doit tenir compte de la façon dont les textes ont été produits, mais aussi de la façon dont ils ont été reçus.

D'autre part, on peut donner à la liste des «genres» ou des «registres» (pour reprendre le terme de Biber, 1995) une extension plus ou moins large, ou plus ou moins limitée. En ce sens, le codage préalable des genres est le domaine qui aujourd'hui encore suscite le plus grand nombre de réflexions (cf. Rastier, 2000). D'autant plus que, les derniers travaux en la matière montrent qu'une typologie rigoureuse devrait s'appuyer aussi bien sur des critères «internes», intralinguistiques (structures et unités) que sur des critères «externes», extralinguistiques (objectifs et situations des textes).

De fait, à l'heure actuelle, si les recherches sur les critères internes n'en sont qu'à leur début,[6] celles qui concernent les critères externes ne cessent de s'affiner.

Sans entrer davantage dans les détails, je terminerai cette première partie de mon exposé en signalant:

[6] Beaucoup de linguistes pensent d'ailleurs que c'est encore de l'ordre de l'utopie, même s'ils reconnaissent qu'il y a eu de grandes avancées dans le domaine.

1) qu'il existe plusieurs corpus constitués selon cet objectif, l'exemple le plus représentatif étant le *British National Corpus* ou encore le Corpus de Référence du Portugais Contemporain, cf. Bacelar Do Nascimento (2000[7])[8]
2) qu'en ce qui concerne le français de France, il n'existe toujours pas de grand corpus de ce type; les grands corpus disponibles sont plutôt des «assemblages», voire des recyclages de données existantes….
   Ou encore des corpus, certes échantillonnés, certes diversifiés, par exemple ceux cités au début: le PFC et le CRFP, mais d'une part, cela ne concerne que l'oral, et d'autre part, ce ne sont pas en tant que tels de vrais corpus de référence…Même celui qui en porte le nom.
3) que la problématique autour de cet objet ne cesse d'évoluer. Pour exemple, les définitions proposées par Rastier (2005) qui distingue 4 niveaux:

   – l'archive (ensemble de données non organisées)
   – le corpus de référence (ensemble de textes sur lequel on va contraster les corpus d'études)
   – le corpus d'étude
   – et le sous-corpus de travail en cours.

Ou encore, les définitions proposées par MacEnery, Xia et Tono (2006) qui distinguent seulement deux niveaux:

– l'archive, pour désigner un collage de productions sans projet défini
– et, le corpus qui suppose cette fois une sélection et des choix préalables.

Autrement dit, on voit que la tendance actuelle est de réserver le terme «corpus» essentiellement pour renvoyer à un objet planifié et construit. Ceci dit, est-il si important pour la description de la langue d'avoir à sa disposition un outil de ce type? Les corpus textuels ne sont-ils pas suffisants? Pour répondre à ces questions, je vais revenir sur deux corpus que je connais bien, celui du GARS et le CRFP, et montrer en quoi ils se sont révélés complémentaires.

3. Le corpus du GARS et le Corpus de Référence du français parlé

Le corpus dit «du GARS» compte environ 2 millions de mots. Ce corpus a été composé au fur et à mesure du développement des recherches; c'est un corpus «ouvert» qui s'est enrichi au fil des années et qui a toutes les caractéristiques d'un corpus «textuel», d'une «archive».

Il est constitué de textes suivis,[9] et comporte des productions orales de types différents (entretiens, conversations, récits de vie, prises de parole publique, explications techniques, enregistrements d'émissions radiophoniques ou de télévision, etc.).

[7] Voir aussi le site: http://www.clul.ul.pt/sectores/projecto.crpc.html, qui donne une présentation complète du Corpus de Référence du Portugais Contemporain.

[8] A mentionner aussi l'importance qu'ont eue les projets européens, et notamment le projet NERC (*Network for European Reference Corpora*), pour la réflexion et la constitution des corpus de référence.

Les locuteurs sont d'origines sociale et géographique différentes, ce sont majoritairement des adultes, mais le corpus comporte aussi un certain nombre d'enregistrements d'enfants et d'adolescents.

Le corpus du GARS n'est pas échantillonné et la répartition entre ces diverses composantes n'est pas équilibrée; il permet cependant grâce aux fiches signalétiques qui introduisent chaque transcription de créer des sous-corpus pour mener à bien des recherches spécifiques.

L'absence de ce souci d'équilibre s'explique en grande partie par les préoccupations premières de l'équipe qui devait avant tout constituer une banque de données orales rendant possible l'étude du «français parlé», pris comme un tout, et contribuer ainsi à la description grammaticale de la langue.

Le lissage des «genres» permettait de créer un objet de description le plus général et le plus neutre possible que l'on pouvait alors comparer plus aisément au «français écrit».

Et, de fait, la prise en compte «globale» de ces données orales a permis d'aller plus loin dans la description de faits déjà connus, de renouveler les études distributionnelles, de préciser les relations entre la morphologie, la syntaxe et le lexique, cf., entre autres, Blanche-Benveniste, Bilger, Rouget et van den Eynde (1990).

Cependant, durant toutes ces années, l'équipe du GARS a accumulé des expériences de classements des productions orales qui n'aboutissaient pas à des «genres», en tant que tels, mais qui permettaient de mettre en corrélation certains types de productions avec certains phénomènes linguistiques.

Ces constatations, mais aussi l'absence d'un corpus échantillonné sur le français, nous a donc amenés à proposer la constitution d'un nouveau corpus basé, cette fois, sur des critères prédéfinis.

C'est donc en 1998, qu'a débuté, comme première étape d'une enquête qu'on aurait voulu plus vaste, la constitution du *Corpus de Référence du Français Parlé*.[10]

L'objectif principal du projet était de pouvoir disposer d'un témoignage de la langue française telle qu'elle se parle aujourd'hui sur l'ensemble de l'hexagone. Il s'agissait avant tout de recueillir des données d'un français parlé que l'on pourrait qualifier d'usage «général et courant».

Pour ce faire, nous avons retenu comme critères d'échantillonnage: la situation géographique, la situation de parole, l'âge et le niveau d'études du locuteur.

---

[9] Les enregistrements qui constituent ce corpus ont deux types de durée; ce que nous appelons «petits corpus» font de 10 à 15 minutes; les «grands corpus» font entre 60 et 90 minutes, ou même plus.

Ce corpus compte aujourd'hui environ **440 000 mots**, il est constitué de **134** enregistrements dont on a transcrit **16 minutes** (en moyenne). La totalité représente près de **37 heures** de parole, et les transcriptions près de 1600 pages. Le son et le texte ont été alignés et le corpus peut être exploité à l'aide du concordancier *Contextes* développé par Jean Véronis (qui permet d'obtenir des concordances sonores, et qui est accessible aujourd'hui par le Net).

Les paramètres pouvant être observés lorsque l'on travaille sur le *CRFP* sont les suivants:

3.1. Variables liées aux locuteurs

– **le sexe**: la répartition par sexe n'a pas été un critère initial mais a pu être constatée
– *a posteriori*; les interviewés se répartissent en 44% de femmes et 56% d'hommes.
– **l'âge et le niveau de scolarité** sont des paramètres qui n'ont pu être vérifiés que pour 80% du corpus (les enregistrements de *parole publique* se prêtant moins à ce type de sélection):
  • 3 tranches d'âge ont été retenues: 18-30 ans (30% du corpus); 30-65 ans (50%); + de 65 ans(20%).
  • 3 niveaux scolaires: collège (29% du corpus); Bac (42%) ; supérieur (29%).
– **la région d'origine**: les données ont été recueillies dans 37 villes de province de dimension moyenne comme Poitiers, Perpignan ou plus grandes comme Lyon et Bordeaux ainsi qu'à Paris; la répartition est la suivante: Paris: 20%; Zone Nord: 37%; Zone Sud: 43%.

3.2 Des variables tenant aux situations de parole

Nous avons défini trois situations d'enregistrement:

– **la parole privée**: ce sont des entretiens sollicités portant en priorité sur des récits de vie ou la présentation d'un savoir-faire. Cela correspond à **63%** des enregistrements.
– **la parole professionnelle**: ce sont des entretiens également sollicités mais dans lesquels le locuteur est enregistré dans l'exercice de sa fonction ou quand il parle de sa profession sur son lieu de travail. Cette situation représente **17%** des productions.
– **la parole publique** : dans tous les cas, le locuteur s'exprime devant un public (réunions politiques ou associatives, émissions radio, etc.). **20%** des enregistrements du corpus sont de ce type.

3.3 Paramètres non utilisés

En revanche, nous n'avons pas retenu comme critère **initial d'échantillonnage** les paramètres liés aux «tâches langagières», ou aux contenus thématiques développés par le locuteur, en ce sens, ces derniers sont beaucoup moins faciles à exploiter.

[10] cf. le numéro 18 de la revue *Recherches Sur le Français Parlé (RSFP)*, pour une présentation détaillée du corpus. Nous en reprendrons les grandes lignes par la suite.

Si on veut en rendre compte, il faut obligatoirement se replonger dans chacune des productions en s'aidant, entre autres des informations fournies dans la fiche signalétique portant sur le contenu de l'enregistrement. Et, dans ce cas, la recherche demande de ne pas se limiter à des extraits de 16 minutes mais d'intervenir sur des subdivisions bien plus précises.

A noter enfin, que ce corpus est moins équilibré que prévu, et qu'il est loin d'être un véritable **corpus de référence**, au sens anglo-saxon du terme.

Pour qu'il le devienne, il faudrait pouvoir le compléter, car trop de situations de parole sont absentes, et trop d'usages de la langue ne sont pas représentés.

Ce corpus comporte donc un certain nombre de lacunes, et sa taille, somme toute modeste, peut poser problème. Néanmoins en dépit de ces défauts, il se présente déjà comme une ressource supplémentaire de grand intérêt, pour parfaire la description du français, et plus encore, pour creuser la piste prometteuse des «genres».

De fait, travailler sur des données, à la fois diversifiées et équilibrées, permet de montrer que les phénomènes grammaticaux peuvent varier de manière significative selon le type de production ou la situation de parole.

Autrement dit, si on tient compte des différents types d'oraux ou des différents types d'écrits, on note que rapidement l'opposition écrit/oral cesse d'être, en tant que telle, pertinente.

Deux petits exemples, en guise d'illustration: un portant sur la forme *contre,* l'autre sur certains emplois de la conjonction *et*.

4. La forme *contre*

Un travail précédent (Bilger et Cappeau, 2003), mené à partir de corpus plus grands de français parlé et de presse écrite, nous avait permis de montrer l'existence d'associations privilégiées de *contre* avec certains noms et certains verbes, ainsi que l'existence de distributions et d'emplois différenciés selon les types de productions. Mais l'étude des seules données du *CRFP* nous a permis de préciser certains résultats antérieurs.

Ainsi, nous avions remarqué que dans le corpus oral, la locution adverbiale *par contre* représentait plus de 53% des occurrences de cette forme, alors qu'elle était quasi inexistante dans le corpus de presse écrite (cela correspondait seulement à 1% des emplois).

Cette dissymétrie s'explique sans aucun doute par la pression de la norme qui porte, depuis longtemps et encore aujourd'hui, un jugement négatif sur la séquence *par contre*.

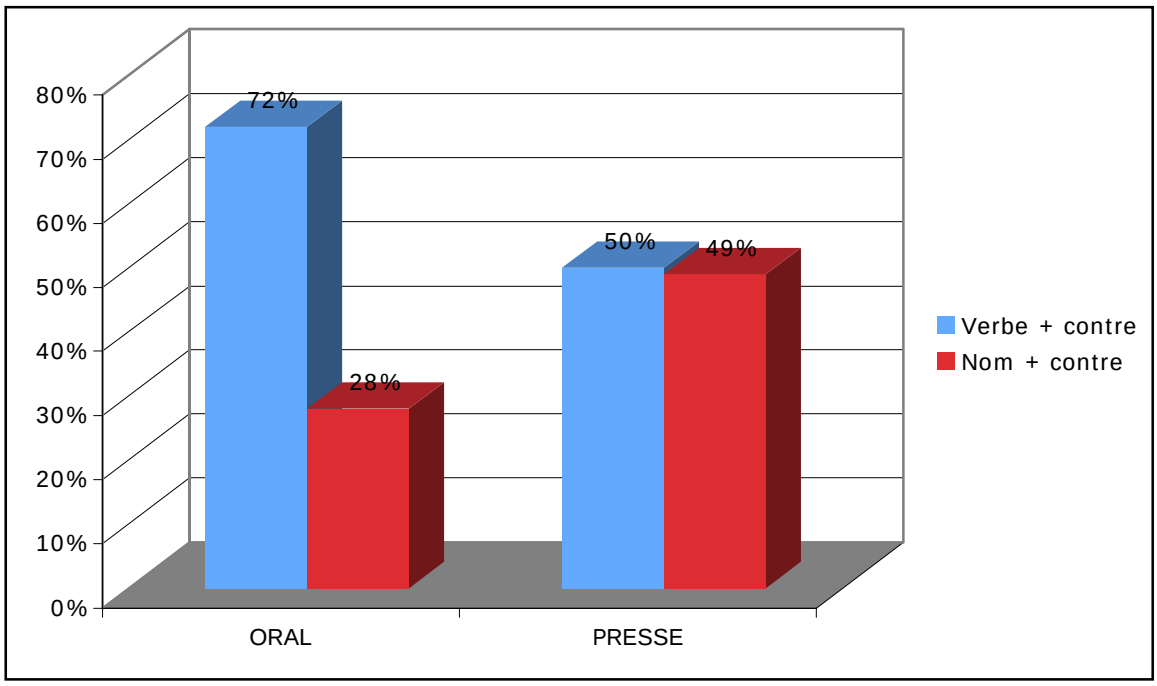
Les données du *CRFP* confirment ces résultats puisque cette locution correspond à 66% des emplois de *contre*. Cependant, si cette locution *par contre* semble effectivement être spécifique des productions orales, il n'en demeure pas moins vrai que seule la 'parole privée' en favorise nettement l'apparition.

Dans ce sous-corpus, *par contre* correspond à plus de 76% des emplois, mais ce taux baisse à 69% dans le corpus 'parole professionnelle' et se réduit à 42% dans celui de 'parole publique'.

En ce qui concerne la répartition des emplois de *contre* dépendant d'un verbe (comme dans *réagir contre, n'avoir rien contre, être contre)* ou dépendant d'un nom (comme dans *la lutte contre, le crime contre, la bataille contre*), nous avions noté, là encore, une disparité entre le corpus de presse et le corpus oral: le corpus oral semblait favoriser les emplois verbaux alors que dans le corpus écrit la répartition entre emplois verbaux et nominaux était équilibrée.

Le Graphique 1 illustre bien cette différence puisque l'on voit que dans le corpus oral, les emplois verbaux correspondent à 72% des occurrences, alors que dans le corpus écrit emplois verbaux et nominaux se partagent de manière tout à fait équitable: 50% et 49%.

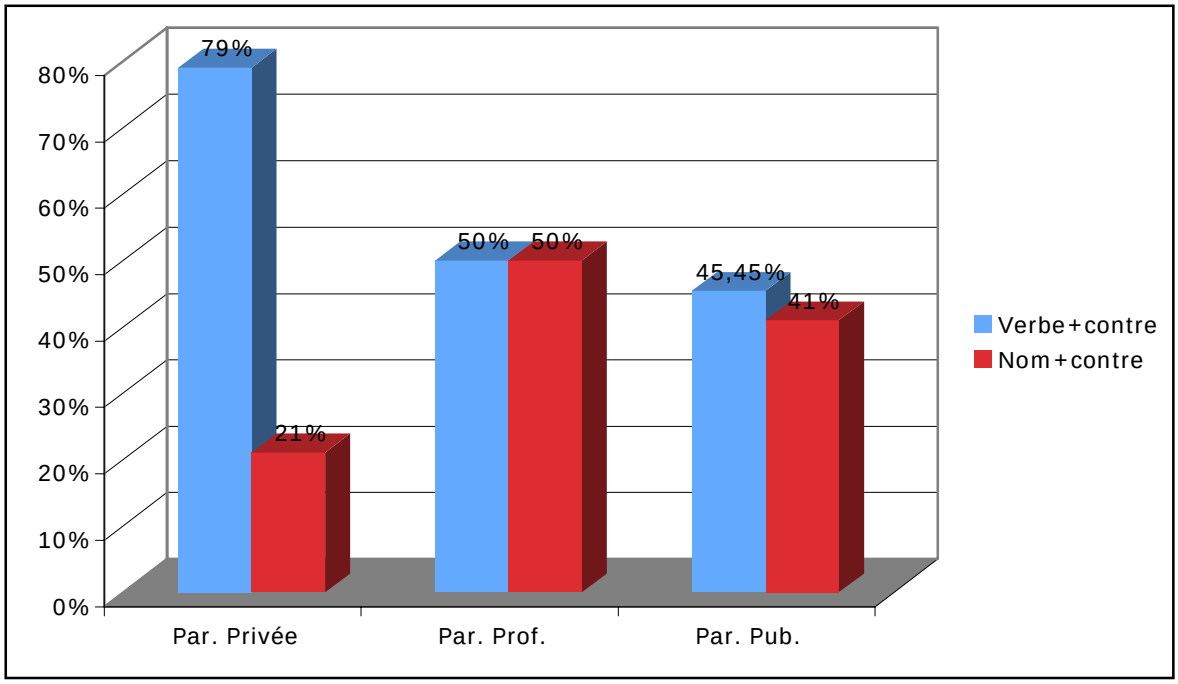
Graphique 1. Les emplois de *contre* dépendants d'une tête verbale ou nominale



Cependant, l'étude des données du *CRFP* permet de préciser là encore que ce décalage ne se vérifie vraiment que dans la situation de parole privée. Dans les deux autres types de situation (parole professionnelle et parole publique), la répartition est tout aussi homogène qu'à l'écrit: cf. Graphique 2, puisqu'on retrouve quasiment les mêmes taux.

Enfin, on avait noté que, dans la presse écrite, l'utilisation de cette forme comme rection locative était rare alors que cela représentait 18% des emplois dans le corpus oral. Cette valeur locative ne semble cependant être réservée, là encore, qu'au seul corpus de 'parole privée'.

Graphique 2. Distinction selon le type de parole



Dans ce corpus, 40% des emplois de *contre* construit par un verbe peuvent s'interpréter de cette manière (par exemple: *être compressé contre les sièges*, *se planter contre les arbres*), alors que dans la parole professionnelle ou publique, *contre* renvoie à un sémantisme en relation avec l'opposition ou l'objection, tout comme dans la presse.

Ces différents résultats montrent donc l'intérêt qu'il y a à travailler sur des données diversifiées. Pour l'écrit, chacun s'accorde à reconnaître qu'il existe des différences importantes selon les genres ou les types. Pour l'oral, cette dimension doit aussi être sollicitée. En ce sens, la simple opposition écrit/oral, même si elle révèle des faits de distribution importants, mérite d'être approfondie et affinée.

5. Le connecteur *et*

Les données de corpus (écrit et oral) montrent l'existence de distributions particulières dans l'usage que font les locuteurs de cette forme. Autrement dit, ce phénomène grammatical de la coordination par *et* est beaucoup plus sensible aux situations de parole et aux genres que ce que l'on pourrait penser.

Ainsi, un travail mené à partir de l'exploitation de 4 corpus de 100 000 mots chacun (un de Littérature, un de Presse, un d'oral professionnel et un d'oral de parole privée), a permis de révéler que ce connecteur est utilisé essentiellement:

– entre deux constructions verbales conjuguées et autonomes:

1) on est vraiment resté de grands amis et pour les enfants ça se passe très bien (corpus oral Professionnel)

– entre deux formes nominales :

2) un ralentissement de la croissance et des importations (corpus écrit Presse)

– ou entre deux formes adjectivales :

3)- cette lettre, confidentielle et personnelle (corpus écrit Presse)

Mais que selon les corpus, ces différents emplois ne se répartissent pas de manière équilibrée:

Voici les proportions pour chacun des corpus (Tableau 1)

Tableau 1. Les différents contextes d'emploi de *et* dans les corpus

	Littérature N = 306	Presse N = 315	Oral Prof N = 329	Oral privé N = 380
et entre constructions verbales	**23%**	8%	**69%**	**68%**
et entre substantifs	**35%**	**58%**	**13%**	**10%**
et entre adjectifs	17%	**14%**	3%	3%
Total	75 %	80 %	85 %	81 %

Les deux corpus oraux se caractérisent par un emploi massif du *et* connecteur de constructions verbales (69% et 68%). Le corpus presse par un emploi également massif mais cette fois du *et* connecteur de séquences nominales (58%). En fait, seul le corpus littéraire exploite les trois contextes à plus de 15%, donc de manière plus équilibrée.

Les résultats obtenus pour l'oral ne sont pas surprenants, ils viennent au contraire conforter certaines analyses antérieures (Halliday, 1985) qui signalent la rareté des adjectifs ou des constructions nominales dans les productions orales au bénéfice des constructions verbales. Et, dans le cas de la coordination, les différents types de «parole» n'ont cette fois aucune incidence sur les répartitions.

En revanche, les résultats pour l'écrit mettent en lumière certaines particularités liées au genre des productions. L'écrit littéraire utilise de fait une palette d'emplois plus ample et moins contrastée que l'écrit journalistique.

Cette spécificité de l'écrit littéraire se révèle également dans un autre emploi du connecteur *et* le cas où la forme sujet est mise en facteur comme dans :

4) Il trouva une tombe fraîche et creusa son scoop. (corpus écrit Pressse)

Les données des corpus, présentées dans le Tableau 2:

Tableau 2. Pourcentage de *et* quand le sujet est en facteur commun

	Litt.	Presse	Oral.Prof.	Oral.privé
sujet en facteur commun	17,32%	10%	2,43%	0,5%
	N =53	N =32	N =8	N =2

montrent que ce type de réalisation est très rare à l'oral, et quasi inexistante dans la *parole privée*. En fait, à l'oral, la forme sujet est régulièrement réitérée, comme dans:

5) ils les nourrissent et ils s'amusent avec eux (corpus Oral Privé)

Ainsi, au vue de ces résultats, la réalisation sans reprise du sujet peut facilement s'interpréter comme une caractéristique de l'écrit. Cependant, dans la mesure où, dans le corpus littéraire, on en relève presque deux fois plus d'exemples que dans le corpus *Presse*, il semble plus prudent de moduler la remarque initiale de la manière suivante:

– la mise en facteur du sujet est une variante stylistique liée à un certain type d'écrit, beaucoup plus qu'une caractéristique de l'écrit en général.

L'ensemble de ces résultats mériterait d'être vérifié sur un plus grand nombre de données, mais il semble que l'on peut déjà les interpréter comme étant révélateurs d'une certaine distribution qui s'opère dans l'usage des locuteurs en fonction de la situation (oral/écrit) et du genre dans lesquels ils évoluent.

La prise en compte d'informations distributionnelles et quantifiées permet de renouveler la présentation que l'on peut donner des phénomènes linguistiques étudiés, et notamment d'éviter de mettre en avant des séquences peu fréquentes, comme cela se produit souvent dans les manuels de grammaire ou dans les dictionnaires. Par exemple, en ce qui concerne la forme *contre*, il peut paraître étonnant que dans les dictionnaires (par ex. *le Petit Robert*, 1990), l'acception première est celle qui renvoie au locatif, alors que celle-ci est de toute évidence la moins utilisée.

Tous ces arguments, qui militent en faveur de ce type d'approche, montrent cependant la nécessité de disposer de corpus qui allient une taille "critique" (c'est-à-dire qui permet de se livrer à des relevés significatifs) et une variété suffisante.

Cette question de variété est d'ailleurs essentielle, puisque, comme nous venons de le voir, c'est ce paramètre-là qui permet de dépasser la frontière entre l'oral et l'écrit, du moins qui oblige à redéfinir l'opposition entre ces deux termes.

De fait, si on accepte de décliner l'oral comme on a l'habitude de décliner l'écrit, c'est-à-dire, en plusieurs types, on note très rapidement qu'il n'y a pas de grandes différences linguistiques entre les deux modes.

Les écarts entre les diverses productions sont à mettre en relation avec des distinctions beaucoup plus fondamentales, comme par exemple, celles de productions «planifiées» et «non planifiées», «formelles/informelles», ou encore celle de situations «soutenues» et «familières».

6. Conclusion

Ainsi, pour en revenir à notre problématique, il va de soi que nous sommes nombreux à regretter l'absence d'un grand corpus de référence pour le français, qui permettrait de mener à bien des études contrastées et beaucoup plus fines sur la langue et ses usages, et qui pourrait être exploité dans les différents domaines relevant des sciences du langage.

Certes, aujourd'hui, on peut dire que le débat sur l'utilité des corpus est terminé ! Tout le monde travaille à partir des données de corpus, autrement dit, tout le monde travaille à partir des données attestées, et c'est sans aucun doute une avancée fondamentale pour notre discipline. Cependant, et sans vouloir remettre en cause l'intérêt de ce mode d'investigation, il est peut-être temps de s'interroger sur l'utilisation même de ces objets, de faire un tri entre les «services» qu'ils rendent et les «mirages» qu'ils peuvent faire naître.

En ce sens, je terminerai par des questions:

– quelle est la portée réelle d'une description basée sur un corpus homogène construit en fonction d'un objectif spécifique? Est-elle autre chose que la description d'un objet clos ? Ou a-t-elle une valeur plus générale?
– quelle est la représentativité d'un usage relevé dans un corpus dont les locuteurs ont été sélectionnés selon des catégories prédéfinies? Et, à quel moment peut-on extrapoler et parler d'un usage général?
– Comment éviter d'avoir une démarche circulaire? c'est-à-dire concevoir un corpus pour vérifier une hypothèse et forcément l'y retrouver?

A ce sujet, il est intéressant de rappeler que l'élaboration d'un corpus basé sur des catégories liées aux locuteurs (comme l'âge, les classes sociales, etc..) est de plus en plus remise en question par les sociolinguistes (cf. Gadet, à paraître), d'autant plus que tout locuteur va pouvoir varier sa production en fonction de la situation de parole, mais aussi, à l'intérieur d'une seule et même production, en fonction de la position qu'il adopte. C'est que nous avions montré dans Bilger et Cappeau (2004) et c'est ce qui avait déjà été signalé dans Miller et Weinert (1998).

Ainsi, il suffit par exemple que le locuteur devienne porte-parole d'un groupe pour qu'il modifie sa façon de parler et qu'il change de registre :

– En ce cas, il utilisera plutôt le pronom *nous* sujet que le pronom *on*, la forme *car* plutôt que la forme *parce que*;

– et le « *ne* » de négation apparaîtra de manière plus fréquente

D'autre part, si ce locuteur est amené à mentionner son activité professionnelle, il risque d'en reproduire les tournures spécifiques et donc, par exemple, utiliser un vocabulaire technique ou encore des nominalisations peu fréquentes dans le langage ordinaire, comme dans les exemples 6) et 7):

6) c'était un médicament contre la toux et un autre médicament qui avait la propriété d'éviter l'expectoration donc contre la toux ça t'empêche de tousser

7) alors toute acquisition dans une bibliothèque municipale est sous la responsabilité du directeur de la bibliothèque mais dans une bibliothèque comme la mienne euh c'est le responsable de la section qui s'occupe des acquisitions

Dans 6, on note que *éviter l'expectoration* renvoie au langage de spécialité mais que ce terme est directement suivi d'une reformulation en langage courant, organisée cette fois autour d'un verbe avec un sujet de type *ça* (*ça t'empêche de tousser)*.

Dans 7, la première construction comporte 2 nominalisations (*acquisition* et *responsabilité*) dont une est en position sujet (*acquisition*) ce qui est rare dans le français parlé de conversation. Dans la deuxième construction, en revanche, le locuteur utilise une façon de dire plus fréquente avec un verbe (*s'occupe*) et une seule nominalisation qui n'est plus en position sujet (*c'est le responsable de la section qui s'occupe des acquisitions*).

Autrement dit, ces exemples montrent l'intérêt qu'il y a à travailler sur des productions suivies, et suffisamment longues, afin de calculer au mieux la variation des usages chez les locuteurs. Cela montre également que la complexité des phénomènes est bien plus grande que ce que l'on pouvait supposer au tout début de la sociolinguistique ou de l'analyse de l'oral.

Les questions soulevées (avant cette petite parenthèse) sont sans doute triviales, mais il me semble qu'il est bon de se les reposer régulièrement: tant il est vrai que le travail du linguiste ne se résume pas à l'exploitation du corpus, mais commence dès la conception de cet outil, et cela nécessite une réflexion préalable.

D'autre part, bon nombre de ces problèmes trouveraient une solution si l'on pouvait disposer d'un point de repère que serait ce grand corpus de référence, tant désiré mais pas encore disponible.

Références

Baude, O. (éd.) (2006) *Corpus oraux. Guide des bonnes pratiques.* CNRS, éditions *Centre de Ressources pour la description de l'oral* (le CRDO).

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: CUP. *Centre de Ressources pour la description de l'oral* (le CRDO).

Biber, D. (1995) *Dimensions of register variation: a cross-linguistic perspective.*

Cambridge: CUP.
Biber, D. Johansson, S., Leech, G., Conrad, S.& Finegan, E. (1999) *Longman grammar of spoken and written English*. Harlow: Pearson Education.
Bilger, M. et Cappeau, P. (2003) Les emplois de *contre* dans les corpus de français parlé et de presse écrite. *Recherches linguistiques*, 26: 91-111.
Bilger, M. et Cappeau, P. (à paraître) De la constitution des corpus oraux à l'analyse: exemples en syntaxe. In: *Interactions et Langages, ICAR*, 1.
Blanche-Benveniste, C, Bilger, M., Rouget, C. et van den Eynde, K. (1990) *Le français parlé. Etudes grammaticales*. Paris: CNRS Editions.
Cappeau, P. et Gadet, F. (2007) Maître-mot et pierre philosophale: l'exploitation sociolinguistique des grands corpus. *RFLA*, vol XII-1: 99-110.
Cappeau, P. et Seijedo, M. (2005) *Inventaire des corpus oraux en langue française*. (http.// www.dglf.culture.gouv.fr)
CRFP: *Corpus de Référence de Français Parlé (DELIC)* (2004) (http://www.up.univ-mrs.fr/delic/crfp), Cf. n°18 de *Recherches Sur le Français Parlé*: Presses Universitaires d'Aix-en-Provence.
Durand, J. et Lyche, C. (2003) Le projet PFC (Phonologie du français contemporain) et sa méthodologie. In: E. Delais-Roussarie and J. Durand (eds.) *Corpus et variation en phonologie du français: méthodes et analyse.* Toulouse: Presses Universitaires du Mirail, disponible à http://www.projet-pfc.net/?pfc-rc:bibiliopfc.
Francard, M., Geron, G. et Wilmet, R. (2002) La banque de données Valibel In: C. Push et W. Raible (éds.) *Romance Corpus Linguistics, corpora and spoken language*. Tübingen: GNV, 71-80.
Gadet, F. (à paraître Compte rendu sur le n° 18 de Recherches Sur le Français Parlé *BSLP*.
Habert, B. (2000) Des corpus représentatifs : de quoi, pour quoi, comment In: M. Bilger (éd.) *Linguistique sur corpus-Etudes et réflexions*. Perpignan: Presses Universitaires de Perpignan, 11-58.
Habert, B., Nazarenko, A. et Salem, A. (1997) *Les linguistiques de corpus*. Paris: Armand Colin.
Halliday, M.A.K. (1985) *Spoken and Written Language*. Cambridge: CUP.
MacEnery, T., Xia, R. et Tono, Y., (2006) *Corpus-based language studies*. New York: Routledge Applied Linguistics.
Martin, E., (1996) Les corpus textuels de l'Inalf.*RFLA* Vol 1-2: 84-87.
Meyer, C-F. **(**2002) *English Corpus Linguistics. An Introduction*. Cambridge: CUP.
Rastier, F. (2000) L'accès aux banques textuelles. Des genres à la Doxa. In: T. Cabré et C. Gelpi (éds.) *Lèxic, corpus i diccionaris.* Barcelone: IULA.
Rastier, F. (2005) Enjeux épistémologiques de la linguistique de corpus. In: G. Williams (éd.). *La linguistique de corpus*, Rennes, PUR. 31-45.
Sinclair, J. (1991) *Corpus, concordance, collocation.* Oxford: OUP.
Sinclair, J. (1996) Preliminary recommendations on corpus typology. Technical report. EAGLES.

Mireille Bilger,
Université de Perpignan-Via-Domitia