RÔLE ET PLACE DES CORPUS EN LINGUISTIQUE :

RÉFLEXIONS INTRODUCTIVES

Damon MAYAFFRE

CNRS UMR « Bases, Corpus et Langage », Nice

Résumé : Si tout le monde conçoit désormais que le corpus est un *observable* nécessaire en linguistique, au moins deux approches se font face pour peut-être se compléter. Pour les uns, le corpus est un *observatoire* d'une théorie a priori, pour les autres, le corpus est un *observé dynamique* qui permet de décrire puis d'élaborer des modèles a posteriori. Théorie et empirie, déduction et induction, linguistique de la langue et linguistique de la parole…, en ce moment, l'épistémologie fondamentale de la discipline se joue et se rejoue, parfois avec naïveté, parfois avec force, dans la réflexion sur les corpus.

Le corpus la notion et l'objet risque d'être victime aujourd'hui en France de son succès. Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence ; plus un linguiste, surtout, qui ne le manipule, le caresse ou le maltraite.

L'année 2004-2005 a connu ainsi nombre de colloques, journées, tables rondes sur le sujet [1]. Le fait que *quasi* simultanément les jeunes chercheurs de Paris et de Toulouse organisent des rencontres autour des corpus montre combien le thème est porteur et représente un avenir pour la linguistique mais à condition de ne pas le galvauder. Boudés pendant un certain temps par une partie de la discipline, les corpus tiennent aujourd'hui une forme de revanche dont on peut se féliciter mais qu'il convient de modérer, d'encadrer, de théoriser.

C'est conscientes de ces enjeux que les journées toulousaines, avec un appel à contribution contraignant, insistent sur la nécessaire réflexion ou pause épistémologique autour du *rôle* et de la *place* des corpus en linguistique, reprenant ainsi l'objectif même de la revue niçoise *Corpus* créée, à cette fin, en 2002.

Dans le numéro inaugural de la revue, Sylvie Mellet nous appelait en effet à réfléchir, en constatant que le corpus était devenu autour des années 2000 « *une médiation consciente* », et sans doute indispensable, « entre le chercheur et le fait linguistique » (Mellet, 2002, p. 9). Aujourd'hui, ce propos fondamental, admis par la plupart, peut être repris dans sa forme littérale pour être décliné en plusieurs interrogations concrètes : de quel(s) corpus parle-t-on ? Pour quelle(s) médiation(s) scientifique(s) ? Et pour quelle(s) linguistique(s) ?

Si tout le monde conçoit désormais que le corpus est un *observable* nécessaire en linguistique, au moins deux approches se font face pour peut-être se compléter. Pour les uns, le corpus est un *observatoire* d'une théorie *a priori*, pour les autres, le corpus est un *observé dynamique* qui permet de décrire puis d'élaborer des modèles *a posteriori*.

Théorie et empirie, déduction et induction, linguistique de la langue et linguistique de la parole, en ce moment, l'épistémologie fondamentale de la discipline se joue et se rejoue, parfois avec naïveté, parfois avec force, dans la réflexion sur les corpus.

Pour ouvrir le débat et ces journées, je me propose de faire un historique rapide de la notion avant d'explorer quelques uns de ses enjeux.

1. Linguistique hors/sans corpus *versus* linguistique sur/de corpus : la détente après la guerre froide

Le traditionnel mouvement dialectique est particulièrement indiqué pour dresser un panorama général de l'objet « corpus » en linguistique ces dernières décennies ; l'inconvénient de ce mouvement étant, comme on le sait, de caricaturer la thèse et l'antithèse pour trancher des positions qui en réalité sont toujours plus nuancées que présentées.

Certains linguistes, le prototype étant les génératives, font (ou ont fait pendant longtemps) l'économie des corpus dans leur pratique scientifique c'est-à-dire l'économie de la confrontation de leur démonstration ou de leur modèle avec des données attestées, recueillies et structurées en corpus. Fondamentalement cela résultait d'une interprétation stricte - trop stricte sans doute - de la dualité saussurienne. Selon la phrase apocryphe du Cours - phrase qui traduit, on le sait aujourd'hui grâce aux travaux de Bouquet (1997, 2005), la pensée de Bally et non celle de Saussure - « la linguistique a pour unique et véritable objet la langue envisagée en elle-même et pour elle-même ». Partant, les linguistes devaient s'intéresser à la langue, au système, à la compétence linguistique. Et par là, le corpus, en tant que recueil d'énoncés effectivement émis, en tant que

recueil de performances individuelles produites par des locuteurs debout dans une réalité sociale et historique déterminée (et déterminante), était non seulement négligé mais d'une certaine manière banni. Les corpus - particulièrement les corpus textuels, nous y reviendrons - relevaient d'une linguistique de la parole qui précisément n'était pas de la linguistique mais de la socio-linguistique, de la psycho-linguistique, ou encore de l'analyse de discours, de la pragmatique, de la stylistique, de la littérature.

Je ne m'étendrai pas sur cette histoire semble-t-il révolue d'une linguistique sans corpus ou hors corpus, notamment du côté des syntacticiens ou morpho-syntacticiens, et particulièrement, comme indiqué, des génératives, mais il faut signaler que Chomsky lui-même déclarait encore en 1999, « corpus linguistics does not exist » (Chomsky, entretien avec Baas Aarts, cité par Rastier, 2005, p. 40). Il convient bien sûr de replacer cette dénégation dans l'ambiance de guerre froide entre théoriciens et descriptivistes que connaissait le monde anglo-saxon. Mais retenons qu'une certaine linguistique fondamentale rejette comme impertinente la confrontation avec les données attestées car celles-ci sont par définition impures du côté du système. En corpus, la grammaire universelle se trouve souillée par la culture, la société, l'humeur ou les pathologies du locuteur, les choix, la sélection de l'analyste, etc. Les corpus de données attestées non seulement ne permettent pas de révéler le système mais le brouillent inévitablement ou le parasitent par divers bruits, le rendant ainsi inaudible au théoricien.

A l'opposé, certains linguistes jugent l'utilisation des corpus obligatoire ; certains, par excès, la jugeant même suffisante. Il peut exister en effet une forme de réduction de la linguistique à l'observation de données réelles ou pire encore au recueillement desdites données. Ce fut la mode des banques de textes, à vocation non pas scientifique d'ailleurs mais patrimoniale. Scheer (2004a et b), dans le troisième numéro de *Corpus*, a décrit les péchés de cette linguistique empirique dominée par le béhaviorisme. Les plus militants des descriptivistes ont poussé en effet loin le principe. Au fond, était déclaré vain de vouloir théoriser la langue. Au fond, il n'y avait plus de système, mais seulement des réalisations multiples, variées, imprévisibles qu'il fallait compiler dans des macro-corpus. Finalement, il n'y avait plus de règles ni de structures mais seulement des entorses aux règles, des exceptions particulières que l'on trouvait dans des corpus oraux ou écrits de plus en plus spécifiques. Seul le relevé –si possible exhaustif– d'énoncés authentiques permettait d'avoir un rendu de l'activité langagière.

Cette tendance est particulièrement représentée en sémantique. Les mots ont peut-être une signification en langue mais ils n'ont de sens et de valeur qu'en contexte. En France, par exemple, c'est le postulat d'une entreprise comme

Frantext ou comme le Trésor de la Langue Française qui en est issu. Le TLF, dictionnaire tout à fait original, ne définit pas les mots à partir d'un sens déjà-là ou construit de manière logique (par l'étymologie par exemple), il entend enregistrer des significations d'usage à partir d'exemples effectivement trouvés dans la littérature française des origines modernes à nos jours. C'est aussi, et là je pourrais en parler longuement, le fondement de la linguistique quantitative et de l'analyse de données textuelles. Selon son grand ancêtre, Guiraud (1960, p. 19), un mot « se définit finalement par la somme de ses emplois ». Partant, aidé de l'outil informatique et des facilités qu'offre l'hypertextualité - l'ordinateur et le support numérique ont joué un rôle décisif -, le chercheur convoque, dans des concordanciers automatiques, toutes les phrases ou tous les paragraphes contenant tel mot du corpus étudié ; plus loin, il compte les occurrences des unités linguistiques du corpus, les trie, les fait contraster, grâce à la statistique textuelle.

Ces pratiques linguistiques - parlons ici des moins caricaturales et des plus récentes - s'appuient fondamentalement, ne le cachons pas, sur une remise en cause de la vision antagonique du couple langue/parole, avec une relecture de Saussure particulièrement vivace depuis la découverte des nouveaux manuscrits, les travaux de Simon Bouquet, le foisonnement de la revue électronique *Texto !*ou les numéros récents des *Cahiers Ferdinand de Saussure*. Il y a en effet actuellement une affirmation existentialiste ou phénoménaliste, qui consiste à refuser la dualité saussurienne comme une dichotomie pour la concevoir comme une dyade, car il ne saurait y avoir d'essence sans existence, de système sans actualisation ou encore, selon les mots que Rastier (2005a) reprend à la tradition aristotélicienne, de puissance sans acte. Pour notre débat, les conséquences sont évidentes : affirmer que « la puissance ne préexiste pas à l'acte » (*ibid.*, p. 33), c'est affirmer qu'il ne peut y avoir une grammaire universelle sans réalisation, de langue sans parole c'est-à-dire, pour finir, de linguistique sans corpus.

Face à ces positions opposées donc, le temps est venu, nous semble-t-il, d'une réflexion de synthèse. Ou plutôt d'une mise au point, car désormais tout le monde utilise le terme corpus et se revendique de lui : « corpus textuel », « corpus sémantique », « corpus en phonologie », « corpus en traductologie », « corpus électronique », « corpus diachronique », « corpus littéraire », « corpus de sciences sociales », « corpus bilingues », « corpus en langues anciennes » pour ne reprendre ici qu'un échantillon des syntagmes trouvés dans la littérature actuelle ou les appels à communication immédiats. Cette mise au point est nécessaire car le risque existe que dans une forme de faux dialogue, plein de malentendus, nous utilisions tous désormais le terme mais dans des acceptions bien différentes. Le risque existe aussi que ceux qui défiaient jusqu'à présent les corpus s'en réclament tout à coup mais pour s'en servir seulement d'alibi : le corpus d'abord nié se trouverait ainsi instrumentalisé ; le risque existe encore

que la *linguistique de corpus* telle qu'elle se définit depuis quelques années dans le monde anglo-saxon (Biber, Conrad & Reppen, 1998 ; Tognini-Bonelli, 2001 ; Aijmer and Altenberg (éd.) 2002) ou en France (Habert, Nazarenko et Salem, 1997 ; Rastier, 2001, 2005a, 2005b), qui s'est appliquée à circonscrire la réflexion et à ériger le corpus en objet, se dilue aujourd'hui et se trouve phagocytée par une linguistique générale qui n'a pas, au départ, les mêmes préoccupations.

Oui, de manière critique, on peut penser dans les termes de l'appel à communication que le corpus est *un effet de mode*. Et la notion ne recouvrira bientôt plus aucune réalité précise si les linguistes en usent seulement comme un sésame obligatoire, porteur conjoncturellement dans notre communauté, et non comme un concept scientifique contraignant.

D'une manière plus positive en revanche, on peut se réjouir que les corpus offrent un biais pour reposer quelques grandes questions épistémologiques - voire métaphysiques - de la discipline. Puisse-t-il que ces journées sinon y répondent définitivement en tout cas fassent avancer la discussion toujours renouvelée. Pour ma part, je voudrais simplement relever les deux problèmes qui semblent les plus fondamentaux, pour essayer de développer un point de vue particulier celui d'un linguiste de corpus et plus précisément d'un chercheur travaillant sur de macro-corpus de discours politiques traités grâce aux rigueurs de *l'analyse de données textuelles* (ADT) assistée par ordinateur.

I) Le corpus et sa nature (ou sa composition) où l'on s'interroge, en filigrane, sur l'objet de la linguistique.

II) Le corpus et la méthode de traitement où l'on interroge l'épistémè de la discipline [2].

2. Le point de vue de la *linguistique de corpus*

La singularité du mot "corpus", en français, ne peut pas nous cacher la pluralité des réalités qu'il désigne. D'évidence, il n'existe pas en linguistique un seul type de corpus mais plusieurs. Cette pluralité trahit d'importantes différences dans les visées et les pratiques de linguistes venus d'horizons différents (phonologie, syntaxe, sémantique, etc.) comme le montreront, tout au long de ces deux jours, les diverses interventions.

De manière hiérarchique, on peut distinguer trois grands niveaux de corpus. (Pour une réflexion plus complète voir Bommier-Pincemin, 1999a et 1999b).

-Les corpus lexicographiques ou sacs de mots dont la grande spécificité et l'incroyable avantage est de pouvoir prétendre à l'exhaustivité. Il n'existe en effet plus de difficulté technique à recueillir et à traiter l'ensemble du dictionnaire ; les corpus lexicographiques peuvent donc non seulement être des corpus clos mais des corpus finis. (Voir, à propos des corpus phonologiques, *Corpus 3*, 2004 et la réflexion de Scheer, 2004a, p. 38 et p. 60-61)

-Les corpus phrastiques de grammairiens ou de syntacticiens dont une des particularités, établie par la pratique, est de pouvoir recueillir des exemples non pas attestés mais forgés, non pas trouvés mais controuvés.

-Enfin les corpus textuels qui ne peuvent aspirer ni à l'exhaustivité ni même à la représentativité et qui concentrent toujours des données attestées puisqu'on ne saurait fabriquer artificiellement un texte pour prétendre en appréhender le sens.

Derrière ces types de corpus, ne l'esquivons pas, se profile la question polémique de l'objet pertinent de la linguistique. Il est évident, qu'au départ, la *linguistique de corpus* au sens de Habert *et al.* (1997) ou Rastier (2005a) considère d'abord les corpus textuels. La linguistique de corpus, *stricto sensu*, repose en effet sur l'affirmation forte, d'une certaine manière subversive, que l'objet du linguiste est le texte. Pour beaucoup de linguistes contemporains en effet, depuis la lecture, longtemps retardée en France, de Bakthine ou Hjelmslev, l'objet accompli d'une linguistique adulte n'est ni le signe ni la phrase (« artefact des grammairiens » (Rastier, 2001, p. 30). Le sens naît du *texte* (et, plus loin encore, du con-texte). Celui-ci doit donc être considéré comme l'unité fondamentale d'une linguistique aboutie. Pour Adam (2001, p. 216), qui cite le Fondateur, la cause est même définitivement entendue :

« Si, comme le dit Saussure, *la langue n'est créée qu'en vue du discours*, la linguistique a non seulement pour objet empirique mais pour objet théorique cette unité de communication-interaction langagière qu'on appelle un TEXTE (ou un DISCOURS)… » (la casse est de l'auteur).

C'est dans le cadre de cette linguistique des grandes unités ou d'une linguistique, *science des textes*, qu'il faut ajouter les réflexions de François Rastier, un des rares auteurs à avoir théorisé, par devant le texte, les corpus (textuels) en linguistique :

« Tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent. » (Rastier, 2001, p. 92)

Ou, pour resituer la réflexion dans le panorama général de la théorie « rastierienne » du *local* et du *global* :

« La détermination du local par le global s'exerce en somme de deux façons, par l'incidence du texte sur ses parties, par l'incidence du corpus sur le texte. » (Rastier, 2001, p. 109)

Et Rastier de conclure :

« le texte est pour une linguistique évoluée l'unité minimale, et le corpus l'ensemble dans lequel cette unité prend son sens » (Rastier, 2005, p. 31)

Dès lors, sans doute peut-on définir hautement le corpus comme *le lieu linguistique où se construit et s'appréhende le sens des textes* (Mayaffre, 2005b). En tout état de cause, on comprend que l'objet non-restreint du linguiste est pour certains auteurs non seulement le texte mais le corpus textuel : il semble qu'à l'aube de journées consacrées au sujet, il fallait rappeler ces pensées qui ne peuvent donner au corpus un *rôle* plus déterminant et une *place* plus centrale dans la linguistique contemporaine.

Je ne creuserai néanmoins pas davantage le point de vue de la *linguistique de corpus* car, poussé à son terme, il viserait peut-être à exclure de la linguistique ironie ! ceux qui prétendaient jusqu'ici, seuls, l'incarner.

Au-delà de la différence de niveaux des corpus lexicographiques, phrastiques et textuels –niveaux, nous l'avons vu, qui interrogent sur l'objet de la linguistique, il faut relever la différence de postures scientifiques que l'on peut tenir face à un corpus, car en elles se joue l'épistémè de la discipline.

De manière schématique, la question est de savoir si, selon les termes utilisés dans l'introduction, le corpus, admis dorénavant par tous, est considéré comme un *observatoire* de quelque chose de transcendant ou bien comme un *observé dynamique*, digne d'intérêt, en lui-même, dans son immanence. Savoir, au fond, si le corpus est une chambre froide d'une théorie *a priori*, ou un observé brûlant, autonome, *réflexif* (Mayaffre, 2002a) –car producteur de sens dans l'organisation particulière qu'il propose du parcours interprétatif– qui débouche sur une théorie ou une connaissance *a posteriori*. Savoir si le corpus sert à révéler un sens qui serait pré-existant ou, fondamentalement, à le construire. En termes plus rapides, la question, pour chacun d'entre nous, au quotidien, est de savoir si l'on se fait une conception documentaire du corpus (recueil d'exemples, base de données, échantillons de langue) ou une conception heuristique. Bref, pour certains, le corpus est un *outil* qui permet de rendre compte d'une réalité transcendante (la Langue ?), d'accéder à un monde déjà-là, d'illustrer une connaissance *a priori*, de "découvrir" un savoir déjà su. Pour d'autres, le corpus est un *objet* vivant de recherche et de connaissance, en lui-même, dont la description débouchera sur des modèles sémantiques à inventer.

Ces questions posent le problème crucial, aussi vieux que la science, de l'induction et de la déduction dans les pratiques linguistiques. En ce qui me concerne, l'option méthodologique choisie pour traiter de grands corpus de textes politiques est à dominante inductive. L'ADT, la *logométrie* (Mayaffre, 2005a), *l'herméneutique numérique* (Mayaffre, 2002b) sont appréciées pour leur dimension heuristique *buttom-up*. Partant d'une description exhaustive et systématique des unités matérielles linguistiques du corpus (lettres, mots, lemmes, co-occurrences) effectivement observées, ces pratiques interrogent le chercheur et organisent le parcours interprétatif d'abord de bas en haut (ce qui n'exclue pas, évidemment, dans un second temps, un nécessaire retour *top down* puis un incessant va-et-vient [3]). On objectera peut-être que des hypothèses de travail « théoriques » auront présidé à la constitution du corpus mais celles-ci n'interviendront pas dans le traitement linguistique de sa matière (Mayaffre, 2005a et 2005b).

Ces questions, surtout, montrent que le terme de corpus, par sa richesse même, peut n'être qu'un cache-misère du débat non résolu entre théoriciens et descriptivistes ou encore de celui voisin mais non recouvrant, évoqué plus haut, entre linguistes de la langue et linguistes de la parole. Le consensus autour de la verbalisation du concept « corpus » ne saurait cacher le divorce autour de sa compréhension. D'une certaine manière, on peut penser que la *linguistique de corpus* –qui penche du côté de la description interprétative, de l'induction et du côté de la parole– a gagné, ces dernières années, une manche en imposant les corpus dans la pratique linguistique mais non la partie dans la mesure où ceux-ci peuvent aujourd'hui n'avoir qu'une fonction validante voire illustrative dans le cadre de démarches qui restent à dominante introspectives. En pareil cas, il s'agit d'une linguistique *ayant recourt* aux corpus. On la qualifiera, au mieux, de linguistique *sur* corpus –le corpus comme support– mais non à proprement parler de linguistique *de* corpus –le corpus comme apport. C'est la distinction que Williams (2005, p. 13) dans l'ouvrage le plus récent sur le sujet, reprend à Tognini-Bonelli (2001) entre les études *corpus-based* et les études *corpus-driven.*

De fait, la frontière semble aujourd'hui déplacée mais sans être abolie. Elle ne sépare plus ceux qui utilisent les corpus et ceux qui ne les utiliseraient pas, mais les linguistes qui se servent des corpus pour valider leur hypothèse et ceux qui les servent pour construire leur savoir.

3. Conclusions

Ces quelques réflexions n'ont d'autre ambition que d'ouvrir le débat. Elles se sont appliquées, sous un angle particulier, à reprendre les principales interrogations de l'appel à contribution de ces journées.

Oui, il existe non seulement une pluralité des corpus que nous avons essayé de structurer en trois paliers, en soulignant les contraintes des corpus textuels, mais aussi une pluralité des linguistiques de corpus comme l'avaient pressentie Habert *et al.* (1997) en mettant le titre de leur ouvrage au pluriel. Cette pluralité, à partir du moment où l'on en prend conscience, doit être ressentie comme une richesse et non comme un handicap.

Oui, il existe aujourd'hui une *mode corpus*, mais dont on peut se servir comme moyen de recharger le débat sur la linguistique en tant que discipline. Et la vivacité de la dispute prouve que n'est toujours pas arrêté l'objet (ou les objets) de la linguistique. Cela ne saurait surprendre pour une science qui n'a, dans sa version moderne, qu'un siècle d'existence.

Oui, surtout, les corpus en tant que « médiation » entre le chercheur et le fait linguistique sont le lieu de confrontation entre la théorie et l'empirie. Ils ne sauraient appartenir exclusivement ni au théoricien ni au descriptiviste pour être précisément un point de rencontre au moment de leur élaboration sur la base d'hypothèses de travail et de leur description sur la foi de l'observation.

Autant d'affirmations provisoires, soumises à débat, que chacune des interventions, j'espère, nourrira, critiquera, approfondira.

NOTES

Cet article a paru dans les Actes des Journées d'Etude Toulousaines JETOU 2005, « Rôle et place des corpus en linguistique », Toulouse, 2005, p. 5-17.

1 Ecole d'été du CNRS « Linguistique de Corpus », Université de Caen, 14-19 juin 2004 ; Journée scientifique « Corpus de Sciences sociales : établissement, numérisation, analyses sémantiques », INALCO, Paris, 8 juin 2005 ; COL'DOC'2005, « Recueil des données en Sciences du langage et constitution de corpus : données, méthodologie, outillage», Paris, 16-17 juin 2005 ; JETOU'2005 « Rôle et place des corpus en linguistique », Toulouse, 1-2 juillet 2005 ; 1ères Journées « Corpus en linguistique et en traductologie », Arras, 28-29 octobre 2005 ; Journées « Corpora et questionnements littéraires » organisées par le Sit@t et Modyco, Paris, 15-16 novembre 2005.

2 Une troisième question, d'apparence plus technique, mériterait d'être frontalement traitée : le corpus et sa forme où l'on s'interrogerait sur les conséquences de l'évolution du support. Les corpus aujourd'hui sont tous numérisés : leur composition, leur manipulation, leur traitement,

leur compréhension s'en trouvent affectés. (Voir par exemple Fabre *et al*., 1998 ou Williams, 2005).

3 Affirmer que la connaissance est le fruit d'un va-et-vient incessant entre la théorie et l'empirie, entre procédés déductifs et procédés inductifs n'est pas contestable. L'idée d'une démarche *hypothético-déductive* constitue une tentative de théorisation de cette intrication et de cette dynamique. Néanmoins, ces affirmations nous semblent trop commodes pour être honnêtes et nous exonèrent aucunement d'indiquer notre posture fondamentale. Un va-et-vient implique certes mouvements et interactions mais un va-et-vient a aussi, nécessairement, un point de départ. Ce point de départ est non seulement déterminant en lui-même pour lancer le mouvement *bottom-up* ou *top-down*, mais il trahit surtout un élan spontané, une préférence trop souvent « inavouée », une posture fondamentale (ou initiale) qui ont toutes les chances de poursuivre le chercheur au quotidien, à l'intérieur même du va-et-vient, tout au long de sa recherche.

BIBLIOGRAPHIE

ADAM, J.-M. 1990. *Eléments de linguistique textuelle*, Bruxelles, Mardaga.

ADAM, J.-M. 1999. *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

ADAM, J.-M. 2001. Discours et interdisciplinarité, *Cahiers Ferdinand de Saussure*, 54, p. 201-218.

AIJMER, B. & ALTENBERG, K. (éd.). 2002. *Advances in Corpus in Corpus Linguistics*, Amsterdam, Rodopi.

AMOSSY, R. (éd.). 2002. *Pragmatique et analyse des textes*, Tel-Aviv, Presses de l'Université de Tel-Aviv.

BIBER, D. 1988. *Variation accross speech and writing*, Cambridge, Cambridge University Press.

BIBER, D. 1995. *Dimensions of Register Variation : A Cross-linguistic Comparison*, Cambridge, Cambridge University Press.

BIBER, D., CONRAD, S. & REPPEN, R. 1998. *Corpus linguistics. Investigating language, Structure and Use*, Cambridge, Cambridge University Press.

BOMMIER-PINCEMIN, B. 1999a. *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de doctorat, Paris IV.

BOMMIER-PINCEMIN, B. 1999b. Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative, in A. Condamines *et al.*(éd.), *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, Actes de l'atelier thématique TALN, p. 26-36.

BOUQUET, S. 1997. *Introduction à la lecture de Saussure,* Paris, Payot.

BOUQUET, S. 2005. Après un siècle, les manuscrits de Saussure reviennent bouleverser la linguistique, *Texto !*

(http://www.revue-texto.net/Saussure/Sur_Saussure/Bouquet_Apres.html).

BRONCKART, J.-P. 1997. *Activité langagière, textes et discours*, Lausanne-Paris, Delachaux et Niestlé.

COMBETTES, B. 1983. *Pour une grammaire textuelle*, Bruxelles, De Boeck-Duculot.

Corpus (2002). « Corpus et recherches linguistiques », 1, numéro coordonné par Sylvie MELLET, 175 p.

DALBERA, J.-Ph. 2002. Le corpus entre données, analyse et théorie, *Corpus*, 1, p. 89-105.

Fabre, C., Habert, B. & Issac, F. 1998. *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*, Paris, InterEditions/Masson.

HABERT, B., NAZARENKO, A. & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Colin.

HJELMSLEV, L. [1943] 1968-1971. *Prolégomènes à une théorie du langage*, Paris, Minuit.

HALLIDAY, M. A. K. & HASAN, R. 1976. *Cohesion in English*, Londres, Longman.

MAYAFFRE, D. 2002a. Les corpus réflexifs : entre architextualité et intertextualité, *Corpus*, 1, p. 51-70.

MAYAFFRE, D. 2002b. L'Herméneutique numérique, *L'Astrolabe. Recherche littéraire et Informatique*, (http://www.uottawa.ca/academic/arts/astrolabe/).

MAYAFFRE, D. 2005a. Analyse du discours politique et logométrie. Point de vue pratique et théorique, *Langage et Société*, (sous presse).

MAYAFFRE, D. 2005b. Les corpus politiques : objet, méthode et contenu, *Corpus*, 4, (sous presse).

MELLET, S. 2002. Corpus et recherches linguistiques : introduction, *Corpus*, 1, p. 5-13.

RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.

RASTIER, F. 2005a. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, PUR, p. 31-45. [En ligne sur *Texto !* (http://www.revue-texto.net/Inedits /Rastier/ Rastier_Enjeux.html)]

RASTIER, F 2005b. Discours et texte (première partie), *Texto !* (http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html).

SCHEER, T. 2004a. Présentation du volume. En quoi la phonologie est vraiment différente, *Corpus*, 3, p. 5-85.

SCHEER, T. 2004b. Le corpus heuristique : un outil qui montre mais ne démontre pas. *Corpus*, 3, p. 153-193.

TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work*, Amsterdam, John Benjamin's Publishing.

VAN DIJK, T. 1984. Texte, in Beaumarchais *et al.* (éd.), *Dictionnaire des littératures de langue française*, Paris, Bordas.

WILLIAMS, G. (éd.) 2005. *La linguistique de corpus*, Rennes, PUR.