

L'idée de corpus linguistique d'après la théorie linguistique du sens commun : hypothèses et perspectives

Georges-Elia SARFATI
Université de la Sorbonne - Paris IV

1. Préalables

Cette contribution a pour objet de penser la notion de corpus linguistique dans la perspective de la théorie linguistique du sens commun. Il s'agit notamment de poser sous forme axiomatique certaines propositions théoriques pour servir à la réélaboration d'un concept cardinal pour le traitement de vastes configurations discursives.

On retiendra quatre propositions¹ :

1. l'introduction en linguistique de la perspective transphrastique est la condition de possibilité d'une Linguistique de corpus ;
2. la Linguistique de corpus présuppose, pour être effective, le passage d'une idée claire des tâches de la linguistique de corpus à la construction ainsi qu'au traitement de corpus effectifs ;
3. la Linguistique de corpus procède en fonction d'options théoriques spécifiques qui sont notamment liées au choix de l'analyste ;
4. tout corpus constitue un artefact – en cela il n'est pas un « donné » – puisqu'il résulte d'une série de choix méthodologiques préalables impliquant un prisme théorique ainsi que la définition d'objectifs analytiques.

1. Pour un état de la question ainsi qu'une problématique approfondie de la notion de corpus linguistique, nous renvoyons le lecteur à (Rastier, 2001).

2. Dynamique discursive et découpe linéaire

À tout prendre, les propositions préalablement formulées constituent à ce jour autant de rappels des positions épistémologiques désormais tenues pour acquises par la plupart des théoriciens et des praticiens de la Linguistique de corpus.

L'existence de banques de données considérables, la mise au point de programmes de traitement informatique de ces mêmes données sont désormais le lot commun d'un domaine des Sciences du langage dont les contours sont de mieux en mieux définis.

Il est toutefois raisonnable de penser que les avancées théoriques permettent également d'infléchir non seulement l'idée que l'on peut se faire de la notion de corpus, mais également la manière d'organiser et de découper de véritables corpus.

L'idée de corpus varie en effet à proportion des conceptions de la discursivité, tandis que la découpe (ainsi que les perspectives d'analyse) du corpus varient à proportion des concepts opératoires qui permettent de caractériser plus ou moins finement les mécanismes de stratification d'un ou de plusieurs ensembles discursifs.

Selon la théorie linguistique du sens commun, une configuration discursive doit toujours être comprise et analysée à partir de deux paramètres :

- d'une part en tant qu'ensemble discursif dont la formation ainsi que l'expression sont fonctions d'une institution de sens spécifique ;
- d'autre part en tant qu'ensemble discursif dynamique, sujet à des variations consécutives à ses « reprises » par différents sites d'énonciation.

Précisons davantage ces deux points.

Une institution de sens (IS) est un dispositif socio-discursif qui a pour finalité la production ainsi que la pérennisation de normes de pensée, d'expression et d'action auprès des acteurs sujets qui en participent (Sarfati, 2008). Une institution de sens, de ce point de vue, coïncide avec un domaine de pratique, c'est-à-dire qu'elle se subdivise en communautés de sens.

Ajoutons qu'une institution de sens (compte tenu des communautés de sens qu'elle intègre) produit son propre sens commun. Rappelons ici

que « le sens commun d'une institution de sens se définit par l'ensemble des manières de dire et des savoirs propres aux membres d'une même communauté de sens » (Sarfati, 2007). Autrement dit, considérée à l'aune de sa seule discursivité, une institution de sens se comprend comme la sémiotisation des normes de sa praxis. Prenons un exemple pour illustrer la distinction ici faite entre institution de sens et communauté(s) de sens. La pratique philosophique constitue une institution de sens telle que nous l'entendons, tandis que les différentes spécialités philosophiques (métaphysique, épistémologie, etc.) constituent autant de communautés de sens. On peut aussi préciser cette distinction en admettant qu'au sein d'une même spécialité peuvent coexister différentes communautés de sens (non pas « la » phénoménologie, mais les courants phénoménologiques, etc.).

Envisagé selon sa dynamique, le sens commun d'une institution de sens est susceptible de divers modes de variations typiques. Ces variations relèvent d'une scansion ternaire dont chaque moment relève d'un ou de sites d'énonciation particulier(s).

Nous avons ainsi distingué (Sarfati, 2008) entre le site d'énonciation instituant du sens commun d'une communauté de sens (canon), le site d'énonciation institué (vulgate), ainsi que le site d'énonciation naturalisant (doxa). Prenons un exemple pour illustrer cette distinction tripartite. Soit le domaine philosophique. On le considérera comme une institution de sens, dès lors où il est envisagé dans la diversité de ses conceptions, des connaissances et des valeurs que cette diversité institue, compte tenu du fait que ces productions s'adressent dans un premier temps à des groupes de sujets définis (auteurs, chercheurs, enseignants, étudiants, amateurs). Dans cette perspective, l'œuvre de Descartes détermine une communauté de sens distincte à l'intérieur de l'institution de sens philosophique. Mais le « cartésianisme » s'appréhende d'abord à partir des textes qui le fondent. Textes « canoniques » s'il en est, dont la transmission (notamment dans le contexte de l'initiation à la philosophie) donne lieu à une « vulgate » (au sens premier). La façon dont le discours cartésien essaime, par le truchement de ses différents relais discursifs donne lieu, à terme, à une « doxa » cartésienne qui n'a, le plus souvent, pas grand chose à voir avec la doctrine philosophique saisie dans sa complexité ni son architecture conceptuelle. Pour autant, cette « doxa » peut se comprendre comme l'ensemble des retombées notionnelles de la pensée de Descartes. Ce qu'il en reste, à ce stade, se fige et

se fixe en stéréotypes ou en dénominations convenues (« le cogito », « je pense donc je suis », « le dualisme cartésien », etc.).

Précisons ici que la scansion canon / vulgate / doxa, outre qu'elle permet de discriminer des états de la discursivité afférente aux communautés de sens d'une même institution de sens, repose sur une série de critères fonctionnels dont la valeur opératoire n'est pas contestable².

C'est en effet grâce à la détermination rigoureuse de ces critères que peuvent être identifiés les frontières ainsi que les points d'articulation spécifiques des états distincts d'une même configuration discursive³.

Moments du sens commun	Topique instituée CANON	Topique transmise VULGATE	Topique naturalisée DOXA
Statut discursif	Exposée	Expliquée	Extrapolée
Régime sémantique	Production sociolectale	Transfert sociolectal	Conversion translectale
Portée déictique	INSTITUANTE	INSTITUÉE	DESTITUÉE
Régime d'hétérogénéité	Hétérogénéité constitutive	Hétérogénéité montrée-marquée	Hétérogénéité montrée-non marquée
Orientation pragmatique	Protensive (futur)	Tensive (présent)	Rétensive (passé/présent)
Degré de réflexivité	Auto-référée	Co-référée	Télé-référée
Type de saisie	Précoce	Médiane	Tardive

Schéma

2. Nous renvoyons le lecteur à la lecture de la thèse de Longhi (2008). À partir d'une grande variété de corpus, politiques aussi bien que littéraires, l'auteur établit la pertinence opératoire de la distinction canon / vulgate / doxa, tout en apportant de nouveaux éclairages sur cette théorie, de manière à en montrer les prolongements notamment pour un traitement informatique des données linguistiques.
3. À propos du critère *Régime sémantique* : le canon est une production sociolectale, même si elle est appelée à régir le collectif (droit, religion, etc.). La vulgate lui impose un changement de réceptacle discursif (d'où le terme de « transfert sociolectal »). L'effet généralisateur de la doxa est effectivement l'occasion d'un détachement des contenus, désormais figés, de sorte qu'ils se trouvent, en apparence du moins, relever de n'importe quel discours (d'où le terme de « conversion translectale »). Le critère *Régime d'hétérogénéité* emprunte aux travaux de Authier-Revuz (1995) ; il s'avère particulièrement pertinent pour étudier les lignes de transformation dynamique d'une topique. Le critère *Orientation pragmatique* distingue, à partir de notions empruntées à Husserl, entre trois visées temporelles. De même, à propos du critère *Degré de réflexivité* : le canon ne fait référence qu'à lui seul, la vulgate tend à se rapporter au canon qu'elle explicite ou qu'elle adapte dans une perspective le plus souvent didactique ; la doxa, affranchie de ce lien de dépendance canonique n'y fait que très lointainement référence, sinon allusion (d'où le néologisme de « télé-référence »).

À partir des distinctions à l'instant posées, nous définirons un *corpus linguistique* à partir de *trois perspectives* :

- **Statut**. Un corpus linguistique coïncide avec le sens commun d'une institution de sens, ou, plus précisément, avec le sens commun d'une ou des communautés(s) de sens afférentes à l'institution de sens considérée ;
- **Découpe**. Un corpus linguistique peut donner lieu à trois types de caractérisations. Il s'agit ici de critères qui permettent de caractériser une institution de sens en synchronie :
 - une caractérisation primaire, induite à partir de l'examen de l'une de ses variations types. Dans cette première perspective, l'analyse portera sur l'examen du mode d'organisation discursive (textuels et génériques) de l'inscription canonique d'une institution de sens ;
 - une caractérisation intermédiaire, engagée à partir de l'examen de son second mode de variation type. Dans ce cas, il s'agit d'examiner les modes d'organisation discursifs de ou des vulgates associées à une institution de sens. les mécanismes de transformations discursifs de la vulgate d'une institution de sens ;
 - une caractérisation tierce, conduite sur la base des retombées naturalisées de sa topique. Cette dernière orientation consiste à rendre compte du mode d'organisation discursif de la doxa d'une institution de sens.
- **Axe temporel**. Un corpus linguistique peut également se prêter à un traitement diachronique dont la finalité sera de rendre compte :
 - du mode de variation de l'un de ses trois moments de variation typique. Il s'agirait d'examiner soit les mécanismes de remaniement ou de transformation discursives du canon, de la vulgate ou de la doxa d'une institution donnée, à partir de l'analyse de l'évolution d'une ou de plusieurs de ses communautés de sens ;
 - du mode de variation de l'ensemble de ses modes de variation. Cette perspective, plus ambitieuse, consiste à se donner pour objet la caractérisation complète de la dynamique globale d'une institution de sens.

Qu'il s'agisse de privilégier l'axe de la découpe synchronique ou de l'analyse diachronique du sens commun d'une institution de sens (à

partir de l'analyse d'une ou de plusieurs de ses communautés de sens), la méthode comparative s'impose comme une nécessité, puisqu'il convient d'établir avec clarté les différentes lignes d'évolution d'une variation de type canonique, vulgaire ou doxique.

Les pistes de recherche préalablement exposées définissent la notion de corpus linguistique dans le cadre du modèle standard de la théorie linguistique du sens commun.

Ces procédures peuvent être affinées, à condition de préciser l'analyse du concept de sens commun d'une institution de sens.

3. Organisation topique et découpe modulaire

Si la première analyse (découpe linéaire) permet de discriminer entre trois moments de manifestation du sens commun (*cf. supra* : 2), il convient maintenant d'examiner à quelles conditions un approfondissement de l'analyse du concept de sens commun permettrait d'ouvrir d'autres perspectives.

Nous avons formulé l'hypothèse selon laquelle le sens commun d'une institution de sens agrège quatre constituants : un constituant cognitif, un constituant gnoséologique, un constituant gnomique, un constituant thymique (Sarfati, à paraître).

Sans préjuger de la répartition ni du degré de développement de chacun de ces constituants, nous avons en outre fait l'hypothèse selon laquelle chaque institution de sens articule de manière spécifique ces quatre constituants selon une économie qui lui est propre.

Rappelons la définition et précisons la fonction de chacun de ces constituants :

- le constituant cognitif sélectionne l'ensemble des formes de rationalité commune (« bons sens ») requises dans un domaine de pratique. L'aspect « cognitif » du sens commun renvoie ici à l'ensemble des représentations sémantiques, des possibles logiques requis par une communauté de sens pour mettre en œuvre son propre discours ;
- le constituant gnoséologique sélectionne l'ensemble des connaissances propres à un domaine de pratique. L'aspect « gnoséolo-

gique » du sens commun d'une communauté de sens est très précisément lié à ce qui la distingue comme pratique du point de vue des compétences que celle-ci véhicule (médecine, philosophie, techniques, etc.) ;

- le constituant gnomique conserve l'ensemble des croyances et opinions inhérentes à un domaine de pratique, y compris des croyances et opinions disparates ou opposées. L'aspect « gnomique » recouvre l'ensemble des opinions admises par un domaine de pratique sans que ces opinions soient nécessairement questionnées par ses acteurs ;
- le constituant thymique caractérise l'orientation axiologique du domaine de pratique considéré. À ce titre il est marqué par la dimension évaluative des relations d'objets distinctives de cette pratique. L'aspect « thymique » caractérise, dans une pratique donnée, ainsi que pour les communautés de sens qui s'y rattachent, les rapports d'évaluation que cette pratique noue avec ses « objets ». Prenons un exemple : la notion de souffrance comme l'épreuve de la souffrance ne sont pas également « valuées » par le domaine médical et le domaine militaire. Dans le premier, l'« objet souffrance » est combattu (la lutte contre la souffrance est un prérequis explicite de la déontologie médicale), dans le second, il est recherché (la guerre consiste à infliger le plus grand nombre de dommages à l'ennemi du moment, en ce sens l'« objet souffrance » y est une norme positive, etc.).

Ce qui permet de distinguer le sens commun d'une institution de sens a trait au différentiel de ses constituants, de même que ce qui permet de distinguer le sens commun de la communauté de sens d'une institution de sens par rapport à une autre communauté de sens de la même institution de sens relève de l'inflexion de chacun de ses modules.

Selon les termes de cette seconde perspective, nous définirons un corpus linguistique à partir de deux paramètres. Partons cependant d'une remarque générale : *Un corpus linguistique est fonction de l'économie topique d'une institution de sens, mais aussi des communautés de sens afférentes à cette institution de sens.* Dans cette perspective, le traitement de la topique d'une institution de sens sera conduit soit dans le cadre d'une analyse synchronique, soit dans le cadre d'une analyse diachronique :

- en synchronie, il s'agira d'identifier l'économie normative de chacun des constituants de la topique d'une institution de sens (ou de l'une de ses communautés de sens) ;
- en diachronie, il sera question d'identifier les transformations de l'économie normative de l'un, de deux, de trois ou des quatre constituants de la topique de l'institution de sens considérée.

Observons que selon les perspectives de la découpe modulaire (en constituants), l'analyse de corpus pourra également être conduite selon des critères d'exigence variables, puisque le traitement pourra privilégier soit l'entier soit l'un des aspects de la topique que l'analyste se donne pour objet.

4. Conciliation des perspectives

Bien entendu, les deux perspectives de traitement ne sauraient être arbitrairement distinguées. Dans la mesure où la théorie linguistique du sens commun dispose (a) d'une théorie des modes de constitution de la normativité d'une institution de sens, ainsi que (b) d'une théorie des constituants (ou modules) typiques d'une institution de sens, il pourra paraître utile sinon opportun – en fonction des fins de l'analyse – d'articuler les deux perspectives afin de sophistiquer l'analyse d'un corpus.

Découpe linéaire	Canon	Vulgate	Doxa
Découpe modulaire	Cognitif / Gnoséologique - gnomique - thymique	Cognitif / Gnoséologique - gnomique - thymique	Cognitif / Gnoséologique - gnomique - thymique

Les deux perspectives préalablement abordées permettent, chacune à leur manière, d'induire une analyse détaillée de la topique d'une IS (et des communautés de sens afférentes à cette IS). Il en résulte deux types de découpe et de traitement du corpus :

- *une découpe linéaire* : celle-ci intéresse l'analyse distincte (synchrone) de chaque moment de variation-type, ou bien l'analyse conjointe de ses trois stades variationnels (panchronie vs diachronie) ;

- *une découpe modulaire* : celle-ci concerne l'analyse du statut de chaque constituant ou module de la topique examinée, soit selon chaque moment de la découpe linéaire (statique), soit selon la prise en vue de seulement deux moments ou bien des trois (dynamique).

Mais il importe de préciser que *les deux découpes peuvent être combinées*, c'est-à-dire menées conjointement selon des choix d'analyse précisément définis (analyse d'un moment variationnel avec ses quatre constituants topiques, analyse d'un constituant topique d'un moment variationnel, analyse des trois moments variationnels avec, comme nous l'avons dit, examen de un, deux, trois ou quatre module(s) de la topique considérée).

Enfin, rappelons ici que la Linguistique de corpus gagnerait à se penser en rapport avec deux activités discursives fondamentales. D'abord une activité descriptive guidée par la recherche des mécanismes de sens structurants de chaque discursivité, ensuite une activité critique guidée par la recherche d'un amendement des dispositifs de normes investis dans les institutions de sens.

Selon nous, ces deux interventions réhabilitent la perspective philologique qui se trouve au principe de toute réflexion linguistique conséquente soucieuse de scientificité.

Références bibliographiques

- AUTHIER-REVUZ, J. (1995). *Ces mots qui ne vont pas de soi, boucles réflexives et non-coïncidence du dire*, Paris, Larousse.
- LONGHI, J. (2008). *Objets discursifs et doxa*, Paris, L'Harmattan, « Sémantiques ».
- RASTIER, F. (2001). *Arts et sciences du texte*, Paris, PUF.
- SARFATI, G.-E. (2007). Note sur le concept de sens commun : essai de caractérisation socio-discursive, *Langage et Société*, n° 119 : 63-80.
- SARFATI, G.-E. (2008). Pragmatique linguistique et normativité : remarques sur les modalités discursives du sens commun, *Langages*, n° 170 : 92-108.
- SARFATI, G.-E. (à paraître). Hermès parmi les loups : sens commun, institutions de sens et doxanalyse, dans DELPHINE, D., HUCHON, M., JAUBERT, A., RINN, M. & SOUTER, O. (éds.), *Au corps du texte : hommage à Georges Molinié*, Paris, Champion, « Bibliothèque de grammaire et de linguistique », 34.