Acta fabula
Revue des parutions
vol. 12, n° 8, Octobre 2011
DOI : https://doi.org/10.58282/acta.6544

La culture par les corpus : qualité & quantité en sémantique de corpus

Julien Longhi

François Rastier, *La mesure et le grain. Sémantique de corpus*, Paris : Honoré Champion, coll. « Lettres numériques », 2011, 280 p., EAN 9782745322302.

La culture par les corpus : qualité & quantité en sémantique de corpus

Julien Longhi

Dix ans après *Art et science du texte*, ouvrage qui a marqué le paysage de la sémantique et de l'analyse des textes, François Rastier, éminent sémanticien, publie un ouvrage qui synthétise ses recherches en sémantique de corpus (la plupart ayant fait l'objet d'articles parus depuis 2001). Ce nouvel ouvrage reprend les thèses principales du précédent, mais étaye davantage encore le propos avec des exemples concrets et très éclairants, et appuie également l'orientation déjà esquissée vers les sciences de la culture. Contrevenant peut-être aux règles de la recension universitaire, je dirai en préambule que la lecture de *La mesure et le grain* fut pour moi très agréable. Plus familier peut être de la terminologie de la sémantique des textes qu'à l'époque de ma lecture de *Art et science du texte*, j'ai trouvé le propos tout aussi ambitieux et stimulant, mais également plus accessible et didactique. Les nombreux exemples participent probablement de ce sentiment. La lecture d'un tel ouvrage peut donner lieu à deux réactions : le relevé systématique des divergences scientifiques[1], dès lors que l'on appartient pas au courant initié par F. Rastier ; ou la saisie de la théorie et de ses enjeux selon le point de vue adopté. Choisissant cette seconde orientation, j'ai pu mesurer à quel point le modèle élaboré par F. Rastier est robuste et cohérent. Il l'est d'autant plus que l'œuvre de transmission menée par l'auteur a conduit à la production de nombreux travaux qui participent du vaste mouvement de totalisation des genres de discours analysés par la sémantique de corpus (comme ceux de E. Bourion ou M. Valette). Ainsi, au-delà des genres issus du discours littéraire, le discours scientifique et philosophique sont abordés, et l'usage de corpus numériques permet d'étendre l'investigation jusqu'au codage des textes (passant de la question des données à celle des métadonnées), au web sémantique, en passant par des sujets tels que la détection de sites racistes ou la néosémie.

1

Généralité du projet de sémantique de corpus

La question du corpus

Constatant que la constitution et l'analyse de corpus sont en passe de modifier les pratiques voire les théories en lettres et sciences sociales, l'ouvrage de F. Rastier entend refléter les rencontres de ses disciplines avec la linguistique de corpus. En effet, avec les corpus numériques, les sciences de la culture trouvent de nouvelles perspectives épistémologiques et méthodologiques, alors même qu'elles se trouvent affrontées à des programmes réductionnistes de naturalisation. Ce combat, déjà ancien pour l'auteur, tient notamment aux objections récurrentes formulées contre la scientificité de ses analyses, qui tiendrait au caractère non repérable des événements. Aussi, à la classique dualité entre induction et déduction dans les disciplines d'observation, le renouvellement méthodologique favorisé par les corpus numériques engage l'auteur à substituer le cycle suivant :

> (i) analyse de la tâche et production des hypothèses ; (ii) constitution d'une archive et sélection d'un corpus de référence ; (ii) élaboration des corpus de travail ; (iii) traitement instrumenté de ces corpus, en contrastant corpus de travail et corpus de référence ; interprétation des résultats et retour aux sources textuelles pour valider l'interprétation (p. 13).

Aussi, le corpus de référence sert de médiation entre la langue historique et la langue fonctionnelle, et les textes qui n'appartiennent plus qu'à la langue entrent dans l'*archive*. Pour l'auteur, un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. De fait, tout regroupement de textes ne mérite pas le nom de corpus.

Philologie et herméneutique

Plus largement, en traitant les corpus, la linguistique renoue avec les textes, donc avec la philologie et avec l'herméneutique : la philologie pour les établir et les documenter, l'herméneutique pour les interpréter, y compris dans leur dimension intertextuelle. La numérisation, bien souvent décriée car elle semble donner une importance excessive à l'informatique, est réévaluée car

> ce que le document perd en stabilité, il le gagne en biais d'interrogation [...] on peut par exemple croiser les résultats de plusieurs méthodes pour faire apparaître de *nouveaux observables*. (p. 19)

Ce projet s'intègre donc à une sémiotique des textes, et l'ambition de F. Rastier est d'« intégrer les facteurs philologiques et herméneutiques dans une théorie néo-saussurienne de la sémiosis, pour articuler les concepts de document, de texte et d'œuvre » (p. 26). Pour cela, il formule un modèle sémiotique du texte qui articule non seulement le contenu et l'expression, mais aussi les pôles du Point de vue (concept herméneutique) et de la Garantie (concept philologique). La méthode proposée distingue donc l'archive qui réunit l'ensemble des documents disponibles pour une tâche de description ou d'application, le corpus de référence qui est constitué par l'ensemble des textes sur lesquels on va contraster les corpus d'étude, le corpus d'étude qui est délimité par les besoins de l'application, et enfin les sous-corpus de travail qui varient selon les phases de l'étude et peuvent ne contenir que des passage pertinents du texte ou des textes étudiés.

Sémantique, interprétation, *doxa*

Contenu et expression, mesure et grain

En évoquant les corpus et non les signes, l'auteur souligne que la langue n'est pas un système de signes, car un signe n'est qu'un *passage* certes réduit, d'un ou plusieurs textes auquel il renvoie. Les corrélations entre plans du contenu et de l'expression sont cruciales pour la sémiotique des textes, car elles permettent d'aborder la question de la sémiosis textuelle :

> Dans cet agenda, la sémantique des textes en corpus met l'accent sur deux complémentarités générales : celle des niveaux de langage ou plans de description (morphologie, syntaxe, sémantique) et celle des paliers d'organisation et de complexité (mot, phrase, texte, intertexte). (p. 48)

Il est notable que les travaux de F. Rastier sont d'une grande importance dans la « médiatisation » des écrits de Saussure, qui ouvrent pour la linguistique des perspectives bien plus ambitieuses que les notes qui constituaient préalablement le *Cours de linguistique générale*.

À cette opposition entre contenu et expression, F. Rastier ajoute celle entre qualité et quantité, telle que formulée dans le titre *La mesure et le grain*. S'appuyant sur la textométrie, il montre que l'on passe de la mesure des mots à la mesure des textes, en mentionnant certaines précautions, car ce qui est mesurable ou fréquent n'est

pas forcément intéressant, les unités rares ou absentes étant parfois tout aussi intéressantes. Cela permet de faire émerger de nouveaux observables, des associations entre éléments qualitatif et quantitatif, des inégalités qualitatives, des associations entre unités sémantiques et unités expressives. Finalement, ce projet permet de dépasser la contradiction entre quantité et qualité, et rassure les chercheurs qui verraient dans le projet de sémantique de corpus une mainmise de l'informatique sur les textes.

La critique envers l'analyse de discours

Proposant de concilier qualité et quantité dans le cadre d'une sémantique de corpus, l'auteur se montre — comme à son habitude — critique envers « l'école française d'Analyse du discours », qui « a fait de la description des idéologies un de ses objectifs » (p. 106) : leur étude échapperait de fait à la sémantique, car la notion de « formation discursive » a été entendue comme position de classe, supposée correspondre à une position de parole. L'analyse pourrait repérer les marques de l'énonciation, mais en définitive ce serait à une autorité politique de dire le vrai sur le sens. D'où la réticence de traiter de l'idéologie comme d'un problème sémantique. Sans entrer dans un long débat, nous poserons ici une réserve au propos très général tenu sur l'analyse de discours française : si F. Rastier semble effectivement parler de certains travaux fondateurs de l'analyse de discours, de nombreux courants composent cette « école », dont certains utilisent des méthodes quantitatives, ou des outils d'objectivation des données afin de ne pas donner le privilège à une parole politique de se prononcer sur la question du sens. Pour l'auteur, il faut dépasser l'opposition entre texte et discours. Il se montre ainsi critique également des perspectives ouvertes par la linguistique textuelle (J.-M. Adam), selon lequel le texte serait un discours privé de son contexte, selon l'équation texte + contexte = discours. Selon lui, cette linguistique estime que l'opposition entre énonciation et énoncé, pourrait être résolue par le repérage dans l'énoncé des « marques de l'énonciation ».

La *doxa*

Pour résoudre les problèmes qui se posent à l'analyse de discours, F. Rastier propose de remplacer le terme d'idéologie par « le terme de *doxa*, en entendant par là l'ensemble des normes sémantiques transgénériques et transdiscursives » (p. 106).

Selon lui, l'intérêt de ce changement terminologique est de redéfinir la *doxa* en termes linguistiques :

> Dans la perspective différentielle, elle se constitue par des oppositions sémantiques ; elle n'est pas « dans les mots » mais « entre les mots », dans leurs relations. Comme ces relations ne sont pas statiques mais dynamiques, il faut caractériser les structures doxales (endoxales et paradoxales) : entre les lexies se placent des seuils évaluatifs, et des parcours génératifs et interprétatifs se déploient dans les zones qu'ils délimitent. (p. 108)

Pour cela, il s'attache à décrire les instances de normativité textuelles, en particulier les genres. Grâce à cette description,

> les « conditions de production » trouvent un autre statut d'intelligibilité, car tout texte oral ou écrit appartient à la strate sémiotique d'une pratique sociale : prescrivant les régimes génétique, mimétique et herméneutique du texte, le genre relie le texte à un discours (politique, juridique, religieux, etc.). (p. 56)

Prenons quelques lignes pour saluer ce travail fort intéressant sur la *doxa*, mais regretter aussi l'absence de références en analyse de discours qui traitent justement spécifiquement de la *doxa*, tel que les travaux de G. E. Sarfati (dont le mémoire d'HDR se trouve sur le site *Texto*! de F. Rastier), ainsi que des contributions à la linguistique du sens commun, telles que celles esquissées dans le n° 170 de la revue *Langages*, Discours et sens commun. Étant à titre personnel auteur d'une thèse intitulée *Les objets discursifs : doxa et évolution des topoï en corpus* (2007), ce manque de dialogue/dialogisme entre les problématisations de la *doxa* me semble préjudiciable dès lors que l'objet dont il est question est à la fois louable (voir notre conclusion) et finalement quelque peu marginal dans le champ des sciences du langage. La question du traitement des idéologies dans le cadre de la *Critical discourse analysis* aurait également pu être mentionnée, même à titre critique.

Les applications prometteuses

Portées pour la linguistique et les sciences du langage

Comme implication interne à la linguistique, la sémantique de corpus permet de transformer la polysémie en problème empirique. En effet, la redéfinition du signe comme passage permet d'appréhender la polysémie par des séries de transformations contextuelles, textuelles et intertextuelles. De même, la création

sémantique d'emplois, la néosémie se fonde sur des rapports contextuels, essentiellement des afférences.

Au niveau de l'analyse des textes, le modèle permet d'éclairer la sémantique des textes théoriques. En effet, en règle générale, l'analyse sémantique des textes théoriques se limite à l'extraction de concepts pour la terminologie et la représentation des connaissances. Dans les textes théoriques, les thèmes correspondent à des concepts et des notions. De la dialogique relèvent notamment l'éthos, la représentation de l'énonciateur et du lecteur, les évaluations. De la dialectique relèvent les démonstrations et argumentations, voire en certain cas les structures narratives ; de la tactique, l'étude des positions des unités sémantiques. Censés représenter une réalité, les textes théoriques exercent une fonction mimétique qui nous importe indépendamment de leur valeur scientifique ou philosophique :

> Une sémantique des textes doit décrire les conditions des impressions référentielles (ou « effets de réel ») suscitées par des structures linguistiques : elles configurent en une « ontologie » locale la sémantique des domaines d'investigation que l'on appelle les objets scientifiques. (p. 169)

Ainsi, les concepts seront vus non comme des représentations mentales mais comme des unités sémantiques : des unités textuelles associées à des traits descriptifs, et « les concepts fonctionnent comme des thèmes dans les discours théoriques, qu'ils soient scientifiques ou philosophiques » (p. 175). Parallèlement, des études sur la polyphonie du texte scientifique permettent des observations intéressantes sur les modes de présentation de l'énonciateur. Cette analyse donne donc une vision très pertinente de genres qui sont peu abordés par les analyses textuelles, mais dont la portée est tout à fait remarquable.

Applications concrètes

Des applications concrètes sont également proposées. Ainsi, avec la sémiotique des sites racistes et la prévention de la xénophobie, l'auteur (dans un travail mené avec M. Valette) pointe l'insuffisance des mots-clés, et également la difficulté de discerner automatiquement les sites racistes des sites antiracistes. Les néologismes sont un bon moyen de corréler des dimensions, domaines ou champs sémantiques. De même, il indique que le vocabulaire des deux positions (raciste et antiraciste) se recoupe en partie mais il existe des différences significatives (Hitler et *bougnoul* se trouvent par exemple plus sur les antiracistes). Si les catégories sémantiques restent invisibles, elles peuvent être véhiculées par des grammèmes (comme pour le nombre par exemple). Ainsi, à l'opposition *Nous* vs *les autres* s'en ajoutent d'autres

qui en découlent, comme le nombre (unique/multiple), la pureté (pur *vs* impur), etc. et la combinatoire de ces catégories reste productive. Les champs sémantiques sont également pertinents, et les plus fréquents sont l'invasion, l'animalité, l'impureté. La stratégie de masquage peut se faire avec surimpression d'autres discours (du politique au scientifique dans le discours négationniste, du politique au religieux, du scientifique au judiciaire, etc.). Enfin, les stratégies discursives comme l'euphémisation, l'allusion ou le cryptage sont également spécifiques des sites racistes.

Autre application concrète, celle en direction de la sémantique du web, opposée au web sémantique. En effet, le programme du web sémantique entend remplacer le « Web des documents » par le « Web des données » et prolonge ainsi le programme classique de la représentation des connaissances. Pour la linguistique, l'enjeu concerne l'amélioration des moteurs de recherche, ce qui n'est pas une moindre tâche. Pour y répondre, l'auteur propose l'alternative de

> moteurs de recherche en plein texte qui tiennent compte des avancées de la sémantique textuelle, notamment : (i) la définition d'unités textuelles non strictement bornées et séquentielles (les passages) ; (ii) l'extension du principe différentiel de la sémantique au contraste de corpus, entre discours, genres et sections de textes ; (iii) l'analyse des genres textuels en zones de pertinence différenciées. (p.226)

L'enjeu est alors la production de connaissances à partir de données massives. La sémantique du web serait alors vue comme une étape de médiation pour constituer une sémiotique comparée. Elle tiendrait compte de la diversité des langues, des discours et des genres, des styles, des inégalités qualitatives au sein des documents, de la diversité sémiotique intrinsèque des documents, de la diversité des tâches, et de la diversité des statuts de fiabilité.

Une conclusion vers l'engagement

Le pari de François Rastier, de « comprendre, expliquer, appliquer pour autant que l'on reconnaisse la dimension herméneutique des sciences » (p. 249) est donc réussi, et ouvre des perspectives prometteuses pour la sémantique de corpus comme pour les disciplines qui la côtoient, car elle leur donne à penser et propose une réflexion stimulante et éclairante. Sa conclusion va dans le sens d'un engagement, et le lecteur que je suis retiendra, peut être avec un certain parti pris, l'importance de l'analyse de la *doxa*, « tâche éminente de la sémantique, [qui] l'invite à déployer toute la vigueur critique propre aux disciplines herméneutiques car, en raison de sa clarté même, faite de toutes les évidences (c'est-à-dire de tous les

préjugés), la doxa resterait, sans une extrême exigence critique, définitivement inaperçue et indescriptible ». (p. 130)

PLAN

AUTEUR

Julien Longhi
Voir ses autres contributions