Corela
Cognition, représentation, langage

HS-15 | 2014
Complémentarité des approches qualitatives et quantitatives dans l'analyse des discours

L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques

Marie-Anne Paveau

Édition électronique
URL : http://journals.openedition.org/corela/3598
DOI : 10.4000/corela.3598
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Marie-Anne Paveau, « L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques », *Corela* [En ligne], HS-15 | 2014, mis en ligne le 15 octobre 2014, consulté le 20 avril 2019. URL : http://journals.openedition.org/corela/3598 ; DOI : 10.4000/corela.3598

L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques

Marie-Anne Paveau

« Corpus linguistics does *not* exist » (Chomsky en 1999)
« La linguistique de corpus prend le langage comme elle le trouve » (Sampson en 1994)

Introduction

1 En sciences du langage, même dans ses dimensions empiriques comme en analyse du discours par exemple, il existe assez peu d'interrogations historiques et épistémologiques sur la question des approches qualitatives et/*vs* quantitatives. En fait, si l'on y regarde de plus près, il n'existe pas de courant ou d'étiquette qui se nommerait « étude qualitative » ou « analyse qualitative », contrairement aux études quantitatives qui sont bien identifiées, comme des études de corpus outillées informatiquement, même s'il y a coexistence de plusieurs appellations et méthodes. Il existe en effet d'assez nombreux ouvrages traitant des méthodes quantitatives, qu'il s'agisse de lexicométrie, plus largement de linguistique de corpus, de statistique textuelle ou de TAL ; on trouve par exemple des bibliographies détaillées de ces approches dans Habert *et al.* 1997, Habert 2005a, et plus récemment dans Mayaffre 2010, ce qui contribue à poser et à définir un domaine identifié comme tel.

2 Il semble donc que le qualitatif soit une sorte d'évidence implicite dans l'histoire de la linguistique, constituant l'approche non marquée en quelque sorte. Le résultat est que les études quantitatives semblent être venues interroger et faire émerger le qualitativisme comme type d'approche, alors qu'il était précédemment acquis comme préconstruit.

3 En histoire et dans les autres sciences humaines, c'est le contraire : il existe de nombreux travaux, et également des manuels sur les deux aspects[1] et un *Dictionnaire des méthodes*

qualitatives en Sciences humaines a même été élaboré (Mucchielli 2004). L'alternative qualitatif/quantitatif reste encore à penser en sciences du langage et les questions posées pourraient être les suivantes : qu'est-ce que l'approche qualitative/quantitative fait à la langue et à ses manifestations empiriques, c'est-à-dire aux matières[2] langagières et par conséquent aux disciplines du texte, du discours et de l'interaction (désormais TDI) ?[3] Qu'est-ce que l'approche qualitative/quantitative fait à la linguistique comme science, toujours dans la perspective empirique TDI, et à la nature de ses savoirs ? Comment se comporte l'alternative qualitatif/quantitative dans l'étude des discours natifs en ligne qui posent à la linguistique le problème de leurs caractéristiques technolangagières et technodiscursives?[4]

4 À partir de ces larges interrogations, je propose d'examiner quatre points pouvant donner lieu à un questionnement épistémologique et permettant de soutenir une perspective réflexive sur les sciences du langage sous leurs approches TDI : la symétrie entre qualitativisme et quantitativisme, la nature du savoir linguistique produit, la prise en compte du contexte et la pertinence de l'une ou l'autre approche sur les corpus de discours natifs en ligne.

1. La paire qualitatif/quantitatif est-elle symétrique ?

5 J'ai déjà partiellement répondu à la question dans l'introduction en soulignant l'évidence non questionnée de l'approche qualitative : la réponse est non, bien sûr, mais j'essaie d'en développer ici les raisons.

1.1. « Les précisions de la linguistique ne sont pas de celles qui s'énoncent par des nombres »

6 Cette déclaration définitoire d'Antoine Meillet et Marcel Cohen dans *Les langues du monde* en 1924, citée par André Cailleux en 1953, explicite parfaitement le présupposé de la linguistique « sans instrument »[5]. Ce que confirme plus de quatre-vingts ans plus tard Benoît Habert, montrant que l'outillage ou l'instrumentation des linguistes remet en cause ce présupposé des origines :

> A part des domaines comme la psycholinguistique, la phonétique et la phonologie, une bonne partie de la recherche linguistique ne suppose pas d'instrument particulier. C'est en tout cas un implicite largement partagé, en particulier depuis les débuts de la grammaire générative. Cette situation se trouve aujourd'hui bousculée par la multiplication, particulièrement sensible depuis une dizaine d'années, avec l'émergence du Web, des outils, des instruments et des ressources modifiant les conditions de constitution d'observables et d'analyse de données en sciences du langage. A côté d'une linguistique sans instrument, s'impose clairement une linguistique « à l'instrument » (Habert 2005b : en ligne).

7 C'est donc l'émergence du quantitatif qui fait émerger la distinction voire l'opposition des méthodes et du coup la question épistémologique. Il ne peut y avoir symétrie, puisque le qualitatif est premier et exclusif, et puisque c'est l'apparition du quantitatif qui fait émerger la paire quantitatif/qualitatif. La distinction est à la fois méthodologique, mobilisant la notion de comptage (le qualitatif fonctionne sans comptage alors que dans le quantitatif, c'est le comptage qui permet d'atteindre des résultats), mais aussi épistémologique : dans les approches qualitatives, l'analyse et l'interprétation se font « à l'œil » ou « à la main » ou même, pourrait-on dire, « à l'humain », alors que dans

l'approche quantitative, l'analyse et l'interprétation se fait « à l'instrument » (Habert 2005b), « au logiciel », « à l'humain augmenté » ou même, j'y reviendrai, « au posthumain ».

1.2. Un exemple de qualitatif revendiqué

8 Il est très difficile de trouver en sciences du langage des exemples de recherches dans lesquelles la méthode qualitative soit explicitement revendiquée : des auteurs désormais classiques de l'analyse du discours et plus largement des disciplines TDI comme Jean-Michel Adam, Jean-Pierre Bronckart, Patrick Charaudeau, Oswald Ducrot, Catherine Kerbrat-Orecchioni, Dominique Maingueneau, Sophie Moirand, Alain Rabatel, Eddy Roulet ou Robert Vion, ne disent rien de leur approche et font du « qualitatif » sans le dire et peut-être sans le savoir, c'est-à-dire sans l'interroger. L'analyse « à la main » ou « à l'humain » est donc le point de départ non marqué ; le modèle TDI est non quantitatif.

9 Voici cependant un exemple où l'analyse qualitative est explicitement revendiquée mais non expliquée, le terme *qualitatif* étant simplement posé en introduction et rappelé en conclusion d'un article de Françoise Boch et Francis Grossmann :

> **Introduction**
> A partir de l'observation d'un corpus d'articles publiés dans le champ de la linguistique, cette étude propose une analyse qualitative et exploratoire des marqueurs lexicaux de la convergence (de type *conformément à nos hypothèses*) et de la divergence (de type *contre toute attente*) par rapport aux attentes de l'auteur-chercheur ou de la communauté scientifique
> **2.2. La construction des attentes comme dispositif argumentatif**
> Dès lors que l'on privilégie une observation « au ras du texte », la notion d'attente est d'abord à concevoir comme un dispositif rhétorique général, que l'on peut définir en termes de stratégies d'écriture.
> **3.1. Pour une approche syntagmatique des marqueurs**
> Un relevé lexico-sémantique des marqueurs révèle que certains termes fonctionnent comme des clignotants dans le repérage et interviennent dans des expressions relativement figées.
> - L'« attente » et ses dérivés, qui peuvent servir à indiquer le caractère attendu (*comme attendu, chose attendue, comme nous nous y attendions, comme on pouvait s'y attendre*), ou à l'inverse, la surprise (*contre toute attente, contrairement à nos attentes, de façon inattendue, un résultat inattendu*) ;
> - la « prévision » (*contrairement à nos prévisions, comme cela était prévisible, ces résultats sont prévisibles*) ;
> - la « convergence » (*conformément à nos hypothèses, nos postulats, nos attentes, aux prévisions*), qui sert la validation positive des attentes
> **Conclusion**
> Ces marques lexicales ne peuvent être rapportées à des modes de construction des savoirs qu'en tenant compte du texte dans son ensemble ; ainsi a-t-on pu observer des modèles de prédiction locaux, qui n'impliquent pas que l'article soit entièrement bâti sur un modèle canonique où l'explicitation des attentes en introduction et la discussion des résultats (en termes de vérification ou non des attentes) reconstruisent une logique linéaire de la découverte scientifique. Non seulement la question du plan de texte est importante, mais la construction des attentes de manière plus générale peut se présenter sous des formes variées. Elles appellent une approche qualitative, à même d'envisager les marques lexicales en lien avec d'autres et au niveau global du texte (Boch *et al.* : 109 et 122 ; gras et italique des auteurs).

10 C'est donc, on le voit, une paire bien asymétrique que l'alternative qualitatif/quantitatif, malgré la paronymie qui unit les termes et la construction binaire de la paire elle-même, qui construisent une représentation symétrique des deux types d'approche. On pourrait presque croire qu'il existe un choix égal entre les deux approches ; en fait le choix ne porte que sur le quantitatif, le qualitatif étant préalablement acquis.

2. La nature du savoir linguistique

11 La construction du savoir linguistique à l'instrument implique que la nature même du savoir élaboré en soit affectée. Les méthodes quantitatives ont en effet des effets épistémologiques qui concernent la question de la vérité de la connaissance, la dimension posthumaine du savoir et sa validité institutionnelle.

2.1. Un savoir à l'œil ou au trompe l'œil ? La question de la connaissance vraie

12 Ce que j'appelle savoir à l'œil, ou à la main, consiste en la production de résultats sans instruments, comme ceux des chercheurs cités plus haut, et d'autres très nombreux travaux en linguistique dans les disciplines TDI actuellement[6]. Les modes d'élaboration des savoirs linguistiques posent la question de leur vérité, et, partant, de leur transmission. J'entends la vérité de manière pragmatiste comme une croyance justifiée et non comme une vérité logique dans la tradition platonicienne : une vérité plus proche de la validité que de l'intangibilité. L'alternative qualitatif/quantitatif pose la question de la vérité des savoirs linguistiques : certains sont-ils plus vrais que d'autres ? et lesquels ? l'outillage quantitatif permet-il une plus grande vérité et validité ?

13 J'essaie d'approfondir cette question sur un premier exemple : en 2001, Dominique Labbé déclare, à partir d'une étude quantitative des corpus des pièces de Corneille et Molière, et en s'appuyant sur la notion de distance intertextuelle, que Molière n'est pas vraiment l'auteur de certaines de ses pièces, écrites par Corneille. Le chercheur résume lui-même l'affaire :

> La distance intertextuelle quantifie les proximités entre plusieurs textes. Elle peut être mesurée grâce à un indice normalisé et à une échelle de la distance. Ces outils peuvent être utilisés pour l'attribution d'auteur. Une application est présentée sur l'un des cas célèbre de la littérature française : Corneille et Molière. Le calcul fait clairement la différence entre les deux œuvres mais il démontre aussi que Corneille a contribué à de nombreux chefs d'œuvre de Molière (Labbé 2001 : 213).

14 Si ce résultat est scientifiquement exact, cela veut dire que mon savoir à l'œil est un savoir au trompe l'œil, un savoir faux. Or mon savoir à l'œil sur Corneille et Molière, c'est celui que l'on m'a transmis à l'école, puis au collège et au lycée, voire au programme de l'agrégation, et celui que j'ai moi-même enseigné dans mes classes de lycée. C'est un savoir mémoriel perceptif, intériorisé, une croyance parfaitement justifiée et validée par l'institution et le canon. Quels seraient les apports de son remplacement par le savoir mesuré scientifiquement ? ou, formulé autrement, quel est l'effet ou l'apport d'un changement de régime de vérité dans un domaine de savoir ?

15 L'examen d'un second exemple pourra permettre d'approfondir la question. C'est celui que donnent Claire Lemercier et Claire Zalc dans *Méthodes quantitatives pour l'historien* :

> La quantification permet, par exemple, d'infirmer le lien fait par les contemporains entre peintres étrangers et avant-gardisme à Paris au début du XXe siècle (Joyeux-Prunel 2007) [...]
> Ces réponses amènent d'autres questions, d'autant plus intéressantes qu'elles renvoient aux représentations des acteurs : pourquoi voient-ils ainsi la peinture étrangère [...] ? le chiffre, « au même titre qu'un fragment de texte ou un qu'un tesson d'amphore », peut jouer le rôle d'indice et orienter une intuition, en plus de ses fonctions de description et de définition des possibles (Grenier 1995) (Lemercier & Zalc 2008 : 17).

16 On voit que ce qui est en jeu, c'est la nature de la connaissance vraie à travers le questionnement du rapport entre les interprétations naturelles, ou mémorielles, ou même doxiques (par exemple dans le cas de la peinture avant-gardiste et des peintres étrangers) et les observations étayées par des analyses scientifiques, de type quantitatif. Plus généralement, ce rapport est celui qui existe entre connaissance intuitive ou perceptive et connaissance rationnelle, entre sens commun et savoir scientifique. Dans *Contre la méthode*, P. Feyerabend décrit « l'argument de la tour », celui qui a été opposé à Galilée pour réfuter le mouvement de la terre. On peut le résumer très brièvement de la façon suivante : la simple observation montre qu'une pierre tombant d'une tour a une trajectoire parfaitement verticale et perpendiculaire à la surface de la terre ; si la terre tournait, la trajectoire serait différente ; donc la terre ne tourne pas. Le lien entre avant-garde et artistes étrangers ressemble un peu à cette affirmation : c'est une connaissance perceptive, qui possède son ordre propre de vérité. Paul Feyerabend explique comment se construisent ces évidences à la vérité non questionnée :

> Les phénomènes semblent parler d'eux-mêmes sans secours extérieur ou connaissance étrangère. Ils *sont* ce que les énoncés associés affirment qu'ils sont. Le langage qu'ils « parlent » est, naturellement, influencé par les croyances des générations antérieures, croyances qui ont été nourries depuis si longtemps qu'elles ne paraissent plus être des principes séparés, mais qu'elles sont entrées dans les termes du langage quotidien, et, après l'entraînement requis, semblent émerger des choses elles-mêmes (Feyerabend 1979 [1975] : 75 ; ital. de l'auteur)

17 L'œil est alors un trompe-l'œil, et la quantification sert alors d'outil de révision des croyances. Elle semble être également un outil de vérité.

18 On peut aussi considérer que le savoir issu des méthodes quantitatives est augmenté, au sens qu'a ce mot en informatique et en robotique. En effet, les outils de quantification informatique constituent des augmentations des compétences humaines d'analyse, et fournissent aux humains une mémoire supplémentaire, qui est externe et quasiment infinie.

2.2. Du savoir augmenté aux culturonomics

19 Le quantitativisme fait partie d'une évolution générale de la connaissance, et de bien d'autres domaines, qui fait actuellement passer l'humanité à la posthumanité. Le savoir instrumenté est posthumain dans la mesure où il augmente la puissance de la connaissance humaine. Il s'agit d'une évolution précisément épistémique, dont la philosophie de la connaissance ne s'est pas encore bien saisie d'ailleurs : les travaux sur la nature ou la validité de la connaissance n'ont pas encore intégré, à ma connaissance, le paramètre de l'augmentation.

20 L'exemple des culturonomics illustre bien ce passage au posthumain. Comme le précisent Michel *et al.*, « "Culturomics" extends the boundaries of rigorous quantitative inquiry to a

wide array of new phenomena spanning the social sciences and the humanities » (2010 : en ligne). L'outil proposé par Google fin 2010, Google N-Grams, est un bon exemple de culturonomics : il s'agit d'un moteur de recherche qui fouille dans les 5,2 millions de livres numérisés par Google et accessibles librement au public, un corpus correspondant à 500 milliards de mots. C'est un outil qui permet de tracer des courbes lexicales sur plusieurs siècles, les livres concernés ayant été publiés de 1500 à nos jours. Les textes sont en six langues, anglais, français, espagnol, allemand, chinois et russe, interrogeables séparément). Comme l'explique Erez Lieberman Aiden, l'un des auteurs de l'outil, « l'objectif est de donner à un enfant de 8 ans la possibilité de parcourir les tendances culturelles de l'histoire telle qu'inscrite dans les livres » (cité par Guillaud 2010 : en ligne). *Culturonomics*, ou, comme son nom l'indique, le moyen de connaître les lois de la culture. À partir de ce corpus librement accessible, le mot apparaît comme un organisme vivant, et le programme permet de retracer les trajectoires culturelles des unités lexicales.

21 Aucun humain ne peut fournir une courbe lexicale sur cinq siècles, 5,2 millions de livres et 500 milliards de mots. Quelle que soit sa validité, ce qui n'est pas ma question ici, l'outil Google N-Grams fournit un exemple concret et opérationnel d'augmentation des capacités de connaissance humaine : le savoir aussi, d'une certaine manière, est devenu cyborg.

2.3. La dimension institutionnelle du savoir

22 Si l'on admet que le savoir n'est pas pure production de l'intellect mais qu'il se constitue également de données matérielles et sociales, dont les données institutionnelles, alors on peut penser que les institutions de savoir (universités, groupes de recherche, revues, instituts, instances d'évaluation, etc.), qui valorisent actuellement les méthodes quantitatives, ont une influence certaine dans la constitution de la connaissance. Une remarque de Claire Lemercier et Claire Zalc dans leur petit ouvrage de 2008, qui concerne l'histoire, est également particulièrement valable en linguistique, dans l'environnement actuel où la recherche est suffixée en *-ex* :

> Depuis 2005, les appels à projets de l'Agence nationale pour la recherche, qui, pour mimer les « sciences dures », ont remis à l'ordre du jour le modèle de la grosse base de données constituée pour elle-même par un personnel précaire, tendant d'ailleurs à faire renaître une certaine forme de traitement quantitatif, avec le risque de retomber dans d'anciennes ornières (Lemercier, Zalc 2008 : 13).

23 On voit donc qu'à la question que je posais plus haut de la vérité et de la validité des savoirs scientifiques, l'institution apporte aussi, sous certains aspects, une réponse.

24 Mais revenons à la manière dont les méthodes qualitatives et quantitatives gèrent la question du contexte en linguistique.

3. La question du contexte

25 C'est à mon sens le point le plus critique dans le questionnement ouvert par la question du rapport entre méthodes quantitatives et qualitatives.

3.1. La longue histoire du contexte dans les disciplines TDI

26 Il existe des synthèses récentes sur la question du contexte et les problèmes qu'elle pose aux différentes linguistiques qui en tiennent compte : un numéro de *Pratiques* est consacré à cette question (Achard-Bayle (éd.) 2006), ainsi qu'un numéro de *Corela* qui étend la problématique textuelle et discursive aux approches cognitives (Achard-Bayle, Paveau (éd.) 2007). Dans *Pratiques*, on peut lire un entretien avec Sophie Moirand qui défend une conception *dynamique* du contexte, calée sur la progression de la recherche et donc sur une certaine forme de prise en compte de l'ensemble de l'environnement :

> Actuellement, [...] je me situe dans une conception dynamique du contexte, ce qui, pour le chercheur en analyse du discours qui travaille sur les données empiriques, repose sur le repérage et la prise en compte d'indices de contextualisation de niveaux divers *au fur et à mesure* de l'avancée du recueil des données et des analyses effectuées (mais pas forcément dans l'ordre de l'énumération qui suit) :
> – le cotexte linguistique, le cotexte linéaire du « fil » intratextuel (le *fil horizontal* du discours ou le fil déroulant de l'interaction et *les reprises* de formes diverses qui s'y manifestent) ;
> – le contexte sémiotique et situationnel (le « hic » et « nunc » de la situation de communication) ;
> – le contexte des séries génériques dans lesquelles s'inscrivent les unités discursives empiriques recueillies à l'intérieur d'un domaine, dans une situation de communication ou sur un type de support ;
> – et enfin le contexte socio-historique qui entre de manière exhibée ou cachée dans les configuration discursives, donc tout ce qui relève de l'intertexte et de l'interdiscours (voire de l'histoire conversationnelle), ainsi que des mémoires collectives et des savoirs partagés, c'est-à-dire de l'histoire... (Moirand 2006 : 43).

27 Les notions de « contexte des séries génériques », qui s'inscrit dans la problématique des genres de discours, de nature sociocognitive pour Moirand, et celle de « contexte socio-historique », s'inscrivant dans l'idée d'une mémoire interdiscursive, signalent en effet une prise en compte de l'environnement un peu plus large et matériel que le « simple » contexte situationnel ou communicationnel. Cette conception me semble encore insuffisante pour refléter une véritable prise en compte des facteurs de l'environnement qui entrent dans les mécanismes de production des énoncés et j'ai adopté pour ma part une position intégrative et non dualiste par rapport aux catégories de l'activité humaine : je pose de fait un continuum entre le langagier et le non- ou extralangagier, ce dernier étant constitué par tous les paramètres qui constituent la réalité, dont le paramètre technologique. Cette dimension, fondamentale actuellement puisque de plus en plus d'énoncés sont produits nativement en ligne, est encore peu intégrée aux définitions du contexte dans les domaines TDI : elle est au mieux intégrée dans l'étude des supports, mais non dans l'étude des éléments langagiers eux-mêmes. Or les discours nativement numériques, en particulier sur les réseaux sociaux, présentent des métissages technolangagiers dont il faut mener l'étude. Pour cela, la définition de l'objet langue admise en sciences du langage doit sans doute être un peu modifiée. J'y reviens en 4.

28 Ces approches du contexte sont adoptées par des chercheurs qui travaillent du côté qualitatif, avec des concepts comme la mémoire discursive, le discours-transverse, les prédiscours, les lignées discursives ou l'environnement cognitif, concepts qui ne sont pas élaborés au sein d'une démarche quantitative, et qui ne pourraient d'ailleurs pas y fonctionner.

3.2. Le contexte en linguistique de corpus

29 Il existe plusieurs conceptions du contexte en linguistique de corpus, sans que la notion même ne soit obligatoirement pensée par les linguistes de corpus. Quand elle l'est, ce qui n'est pas général, elle apparaît sous les deux versions du corpus réflexif et du corpus comme contexte.

30 Pour Jacques Guilhaumou, un corpus doit être coconstruit, ce qui le rend réflexif. C'est dans la réflexivité que gît la contextualité puisque l'environnement du chercheur se trouve intégré dans le processus d'élaboration du corpus :

> § 59 Nous constatons donc que l'aspect réflexif du discours prend une importance toujours plus grande dans la procédure de formation de nouveaux corpus. Il s'agit en effet d'associer dans un corpus issu de l'activité langagière de locuteurs précis non seulement des acteurs stratégiques et des acteurs émergents, des experts et des militants ordinaires, des protagonistes et des spectateurs, mais aussi les chercheurs en sciences sociales eux-mêmes. Nous entrons ainsi dans un processus de *co-construction des corpus* où la question de l'articulation du discours et du métadiscours constitue une dynamique interne à la formation de nouveaux corpus.
> § 61 La réflexivité du discours, thème désormais majeur de l'analyse de discours, a ouvert la voie au corpus réflexif non seulement sous la forme du « très grand corpus » où texte et contexte se retrouvent dans un même espace construit, mais aussi sous la forme de corpus co-construits à l'intérieur desquels le point de vue du chercheur est l'une des données majeures de la construction de ces corpus (Guilhaumou 2002 : § 59 et 61).

31 Pour Damon Mayaffre, c'est le corpus lui-même qui est le contexte, et le chercheur va jusqu'à faire des deux notions des synonymes :

> Finalement, la linguistique de corpus n'est rien d'autre qu'une *linguistique endogène*, et son outillage par la statistique (endogène par essence) apparaît naturel aussi bien chez [Biber 1988, 1995, 1998], [Habert, Nazarenko et Salem 1997] que [Malrieu et Rastier 2001].
> Matrice du sens, le corpus entretient, dès lors, un dialogue direct avec la notion de *co(n)texte* puisque l'on admet que la co(n)textualisation est la condition de la maïeutique du sens. Le sens naît en/du co(n)texte, avons-nous plusieurs fois écrit en soulignant la vanité d'une linguistique ou simplement de pratiques méthodologiques décontextualisantes. Le sens naît en/du corpus pourrait-on renchérir liant corpus et contexte dans une relation étroite, quasi synonymique.
> Le corpus peut être en effet conçu comme une forme privilégiée du contexte. Plus précisément, nous définissons le corpus comme *la forme maximale du contexte*. De la lettre au mot, du mot à la phrase, de la phrase au paragraphe ou à la partie, de la partie au texte, du texte au corpus : le phénomène de co(n)textualisation (ou, autrement dit, l'extension de l'objet du linguiste) semble devoir s'arrêter là.
> Précisons cependant : en définissant le corpus comme forme maximale du contexte, nous entendons forme maximale *formalisable* du contexte, car le contexte (le co-texte proche ou l'intertexte plus lointain) est insondable et à proprement parler *insaisissable*. Plus loin encore, le contexte lorsqu'il s'étend au-delà du co-texte ou de l'intertexte pour toucher à la situation socio-historique et aux conditions de production des discours est une chose qui échappe pour partie aux études strictement linguistiques (Mayaffre 2010 : 13 ; ital. de l'auteur).

32 Les précisions sur la formalisation sont essentielles et formulent bien la ligne de fracture entre linguistes du quantitatif et penseurs du qualitatif : dans les disciplines TDI, le contexte est, au contraire de ce qu'affirme Damon Mayaffre, définissable et saisissable ; son existence, sa saisie et sa description fondent même ces disciplines. Évidemment, tout

dépend de l'approche que l'on en forme, et les propositions de Sophie Moirand sur la distinction à faire entre corpus de travail et corpus de référence est justement faite pour régler cette insaisissabilité du contexte global des énoncés. Ma propre approche à partir de l'environnement vise également à proposer un dispositif de saisie de l'environnement, qui neutralise l'impossibilité de la saisie totale, les contextes des énoncés étant quasiment infinis dans le temps et les espaces de la production verbale.

33 Là encore, l'appel aux analyses des historiens s'avère féconde pour la linguistique. La critique qu'ont adressée les microhistoriens aux historiens des grands corpus met par exemple en exergue des points mobilisables dans une réflexion de type linguistique :

> Les microhistoriens ne choisissent pas nécessairement d'abandonner formalisation et modélisation. Leur démarche relève plus du déplacement que de la déconstruction. Ce qui est critiqué, est un travail de quantification utilisé pour mettre au jour des structures préexistantes, de fait déjà connues : celui qui emploie les outils surdimensionnés pour démontrer l'évidence. [...] La démarche est rejetée au profit d'une reconstruction du passé à partir de traces ou d'indices. Carlo Ginzburg en appelle ainsi à passer du « paradigme galiléen » au « paradigme indiciaire » [...]. Il refuse le mythe d'une histoire scientifique, où le tableau de chiffres est en soi une conclusion pour redonner une dignité à l'interprétation de traces » (Lemercier, Zalc 2008 : 15-16).

34 Une notion comme celle d'« exceptionnel normal », proposée par l'historien Edoardo Grendi en 1977, constitue une réponse à ce refus d'une histoire que seule la quantification rendrait scientifique : on postule une certaine représentativité de l'exceptionnel, à titre heuristique, certes, et sans illusion sur la généralisation du phénomène ou de l'individu. « Mais, s'étonnent Claire Lemercier et Claire Zalc, comment savoir si un individu est exceptionnel avant d'avoir évalué les grands traits du groupe dont nous l'extrayons ? » (2008 : 17). Cette évaluation passerait par la quantification. Mais est-ce bien certain ? n'existe-t-il pas une alternative à la paire binaire qualitatif/quantitatif ? le chercheur en sciences humaines et sociales est-il coincé dans ce choix à deux entrées ?

35 Il me semble que certains corpus, comme celui des discours natifs en ligne, se prêtent mal à une approche qui devrait choisir entre qualitatif et quantitatif : la matérialité technolangagière des énoncés en particulier, résiste tant à la quantification qu'à une approche qualitative classique. Pour fournir des observations et des analyses des productions verbales dans les univers discursifs numériques, il faut peut-être proposer des perspectives et des outils qui dépassent cette alternative.

4. Matières technolangagières

36 Dans les univers discursifs numériques, la normalisation des corpus en vue de leur quantification n'est pas pensable : le langagier est traversé par la technique, les énoncés sont délinéarisés, les textes sont hypertextuels et les énonciateurs ubiquitaires. L'approche qualitative classique se heurte aussi aux limites du technolangagier puisque la linguistique n'est pas outillée pour penser une langue métissée de technique.

4.1. Une perspective symétrique pour le discours

37 Pour essayer de penser au delà de l'alternative qualitatif/quantitatif, je propose une approche symétrique des discours (Paveau 2009, 2012c, 2013). J'admets que le discours est une matière langagière constituée par son intégration dans les environnements quelle

que soit leur nature. Il est constitué d'un ensemble de productions verbales élaborées dans un environnement dont les paramètres sont à la fois humains et non humains[7], explicites, implicites et tacites (les prédiscours) et qui ne peuvent en être isolés : l'environnement de production des énoncés est constitutif de leur description, leur fonctionnement et leur sens. Dans cette perspective, le chercheur en linguistique doit éviter le logocentrisme (prise en compte des seuls énoncés) et « l'égocéphalocentrisme » (approche centrée sur le locuteur comme producteur d'énoncés). L'étude est non centrée, l'environnement tout entier étant pris comme objet d'analyse. Cet environnement est mixte : toutes ses données entrent en ligne de compte dans l'élaboration des discours, le social, le culturel, le politique, l'esthétique, l'éthique, le biologique, le technique, l'objectal, le matériel, etc. Les catégories sont pensées comme ouvertes les unes sur les autres, et non closes à l'intérieur de leurs frontières. Dans cette perspective, le langagier n'est plus homogène, mais hétérogène, intrinsèquement mélangé de technique par exemple quand il est produit en ligne. Avant de montrer en quoi les écritures natives en ligne bousculent l'ordonnancement binaire du quantitatif et du qualitatif, je précise les différents statuts numériques des énoncés traités par des programmes ou d'autres outils informatiques et je décris ce que j'appelle les matérialités technolangagières.

4.2. Typologie des écritures numériques et productions technolangagières

38 Le langage, non plus étudié, mais produit « à l'instrument », acquiert le trait « numérique », qui recouvre cependant plusieurs processus :

39 – Le numérique 1 correspond à des textes produits hors programmes informatiques qui seront numérisés pour l'analyse. C'est le sens le plus fréquent de *numérique* dans la linguistique de corpus. La numérisation à partir de textes imprimés linéaires les rend navigables. Le texte ainsi devenu numérique peut se trouver en ligne ou hors ligne.

40 – Le numérique 2 correspond à des textes produits hors programmes informatiques qui sont scannés (non ou peu navigables) ou numérisés (navigables) pour la lecture en ligne ; c'est le cas par exemple des revues mises en ligne a posteriori.

41 – Le numérique 3 correspond à des discours produits nativement en ligne, qui relèvent de deux catégories : ils suivent une écriture linéaire, c'est-à-dire sans liens hypertextuels ni intégration de médias et hypermédias (textes de blogs littéraires ou articles scientifiques en ligne par exemple) ; mais leur écriture peut également être délinéarisée par des liens, des hypermédias, une poly-énonciation (commentaires, wikis), et un ensemble de procédés d'écriture en ligne qui affecte la mise en forme du langage.

42 Le numérique 3 est très peu étudié en sciences du langage quel que soit le domaine et l'écriture en ligne est par ailleurs très peu pratiquée par les linguistes (Paveau 2012d). En revanche, ce type d'écriture est observé par les sciences de l'information et de la communication et, dans une moindre mesure, par les humanités numériques, sans que ces disciplines n'aient cependant beaucoup développé de réflexion sur les formes langagières et leurs traits technolangagiers.

43 Un trait technolangagier est une caractéristique de nature composite due à la production du langage dans un dispositif d'écriture numérique. On peut citer bien évidemment l'hyperlien, qui est un mot ou un segment technolangagier dans la mesure où il est cliquable et support d'hypertexte, le hashtag sur le réseau Twitter (Paveau 2012a), qui est

un technomorphème dans la mesure où il permet par clic de redocumenter un thème ou une information, ou encore les boutons de partage d'information, qui sont des formes technodiscursives de discours rapportés (Paveau 2011, 2012a et b). Certains genres de discours sont des genres technodiscursifs : la demande d'amitié sur Facebook par exemple, parce qu'elle passe nécessairement par le bouton « ajouter », et qu'elle peut de ce fait se passer d'un discours « traditionnel » (un message accompagnant la demande) constitue un genre technodiscursif (Paveau 2012b).

4.3. Pour un traitement écologique des productions technolangagières

44 La « boite à outil » des sciences du langage et de la théorie du discours n'est pas forcément équipée pour rendre compte de ce type de matérialités. La question qualitatif/quantitatif se dissout sous des interrogations qui touchent à l'équipement global de la linguistique pour penser la technodiscursivité des écritures natives en ligne. Le traitement quantitatif, qui impliquerait une extraction des énoncés de leur écosystème numérique, ramènerait le linguiste vers le logocentrisme ou l'égocentrisme. Quant au traitement qualitatif, il reste inadapté faute d'outils et la distinction binaire entre langagier et non-langagier est souvent maintenue, les énoncés du web étant dits être produits *sur* un support technologique et non dans un univers technologique. Dans les deux cas, les linguistes continuent de considérer les énoncés en ligne comme des énoncés purement langagiers, qu'ils relèvent, extraient et alignent en corpus en abandonnant l'environnement numérique, pourtant facilement présentable par de simples captures d'écran.

45 Les travaux qui existent sur les énoncés du réseau Facebook par exemple, procèdent ainsi par extraction et normalisation : dans un travail présenté au cours d'un colloque sur les énoncés trilingues de comptes Facebook de locuteurs algériens (Aït Saïdi, Matthey 2012), les chercheuses restituent les statuts en essayant de conserver les caractéristiques technoscripturales par un jeu de gras et d'italique :

Amel (29 ans, Boston)

(Commentaires à propos d'une photographie de l'enfant d'Amel)

Nezline: Ayouuuuuuuu dak ziiiine !!! khmous khmous khmous !!!!!!!

[ar. ayu: dak ezine]	*[ar/kab khmus khmus khmus]*
Ô cette beauté	cinq cinq cinq
TR: Ô quelle beauté	(formule pour éloigner le mauvais oeil)

Salim: salut cher cousine cava waw trés jolie Amel c'est une vrais poupée allah ibarek, khemouse a3liha, oui c'est vrais elle le regard de Kenza, et un trés jolie sourir machallah, rabi yahfadha inchallah

ar. Allah ibarek	*Khmus a3liha*[3]	*ar. macha Allah*	*ar.rebi yehfedh'ha incha Allah*
Allah béni	cinq sur elle	Ce que veut Allah	dieu la garde s'il le veut

TR: Qu'Allah la bénisse (formule pour éloigner le mauvais œil) telle est la volonté d'Allah, Dieu la garde s'il le veut

Amel: merciiii a tousssss

Rym: oh la poupée machallah elle grandie vite que dieu te la garde, bisou

Amel: lakouba lik!!

[ar. la3quba lik]

Souhait pour toi

TR: Je te le souhaite en retour (i.e d'avoir une aussi belle enfant).

3 3 est mis pour le graphème ع

3

ILLUSTRATION 1. EXTRAIT DE CORPUS DE AÏT SAÏDI, MATTHEY 2012

46 Or, un compte Facebook, c'est de la couleur, des colonnes, des photos de profil, des images, des liens, des dates, des icones, etc., tous éléments qui contextualisent très fortement les productions verbales :

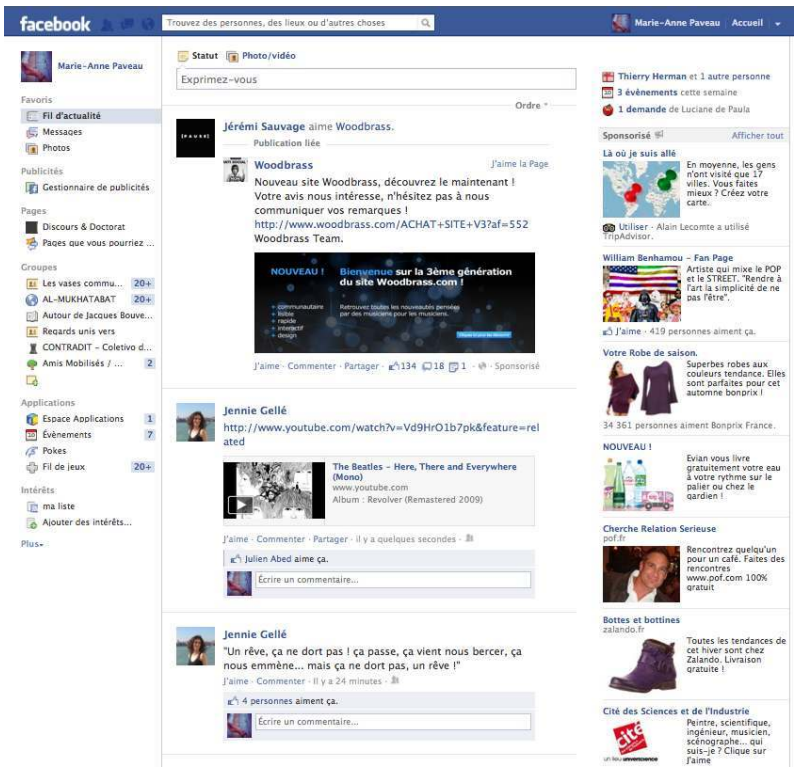

ILLUSTRATION 2. EXTRAIT DE LA TIMELINE DU COMPTE FACEBOOK DE M.-A. PAVEAU AU 10 NOVEMBRE 2012

47 Autre exemple, le réseau de microblogging Twitter. Il serait dépourvu de sens d'analyser les tweets hors du contexte de la timeline, qui montre bien les spécificités technolangagières du hashtag par exemple (mot rendu cliquables par le croisillon qui le précède), et de bien d'autres technomots et technodiscours spécifiques aux univers discursifs numériques :

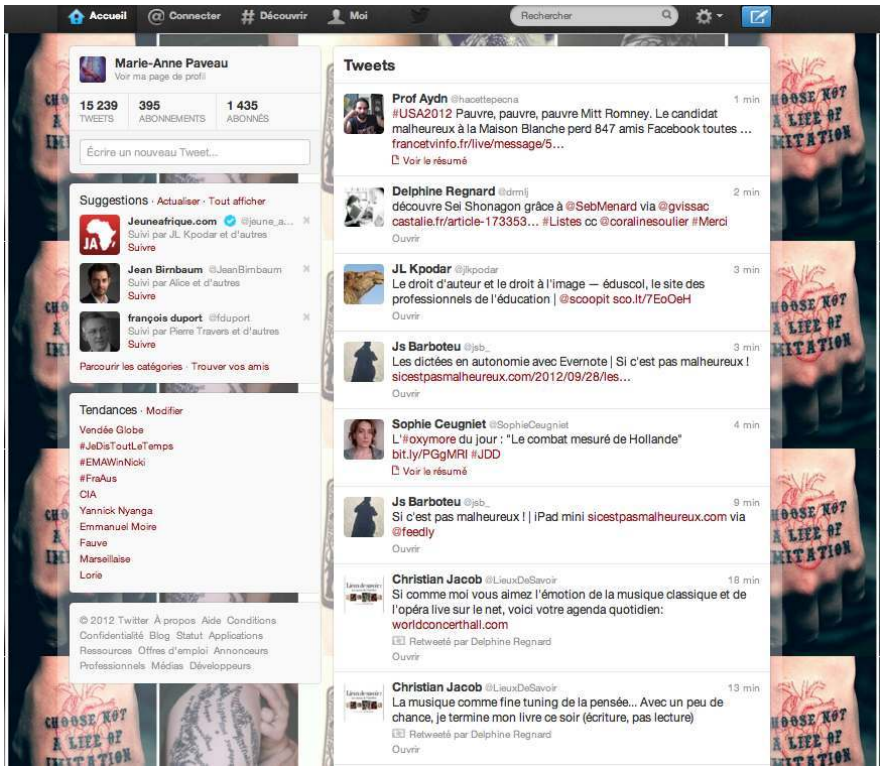

ILLUSTRATION 2. EXTRAIT DE LA TIMELINE DU COMPTE TWITTER DE M.-A. PAVEAU AU 10 NOVEMBRE 2012

Conclusion

48 Alors que le Web 2.0 reste peu étudié par les analystes du discours français, le Web 3.0 se profile déjà : c'est un web de données (data), qui modifie nos représentations du texte et du discours. Les données, un peu rapidement nommées « données brutes », constituent désormais une bonne partie des productions verbales sur l'internet : quasiment tous les états et toutes les collectivités ont leur site d'open data, qui contient un nombre important de données publiques. Les data sont-ils des discours ? Pierre Mounier avait annoncé la « mort du discours » en 2011, au profit des données justement, mais aussi des documents[8] (Mounier 2011). Certes, les données ne sont pas mises en discours de la même manière que les productions verbales inscrites dans la pensée, l'interaction, l'énonciation. Il n'en reste pas moins que la mise en listes ou en tableaux des data sont une forme de discursivité et que, surtout, le phénomène de la redocumentation[9] des données en constitue une véritable remise en discours.

49 Il y a là un enjeu important pour les linguistes, qui sont pour l'instant absents de ces terrains : les formes discursives natives, les formes spécifiques des data, les processus numériques de redocumentation, voilà autant de corpus possibles pour la linguistique, à condition qu'elle sorte de l'alternative qualitatif/quantitatif, et qu'elle travaille sur la

modification de ses objets. Quel sens y aurait-il en effet à quantifier les énoncés du web ? Et surtout comment, dans cet infini discursif, « découper » des corpus, même immenses comme celui que propose Google N-Grams, corpus qui soient justifiés par la raison scientifique ? Il faut sans doute travailler autrement.

50 Dans un texte important de la réflexion sur le numérique en sciences de l'information et de la communication (Pédauque 2003), des chercheurs déclarent que la linguistique textuelle, nommément celle que Jean-Michel Adam a mise en place, est impuissante à saisir ces nouvelles formes de discours que sont les données et les documents. Le verdict est assez rude, mais je pense qu'il est exact. Jean-Michel Adam, cité parce qu'il est le plus productif et le plus prolifique des textualistes, et également, il faut le dire, celui qui produit la théorie textuelle, n'a pas intégré d'interrogation sur le texte numérique ou les dimensions technolangagières des textes nativement produits en ligne. C'est une dimension qui reste étrangère à son travail. Invité en 2006 aux JADT à Besançon (Journées internationales d'analyse statistique des données textuelles), il se donne comme objectif de décrire les points de contact et de partage entre les deux disciplines, qu'il appelle « la linguistique » et « l'informatique », et utilise des données fournies par la numérisation (essentiellement les collocations) pour montrer le bien-fondé de cette interdisciplinarité. Il reste que les textualistes de cette tradition continuent de travailler sur des textes, et non des technotextes, c'est-à-dire des productions natives en ligne, délinéarisées et parfois multimédiatiques. Je n'ai pas cité jusqu'à présent François Rastier qui travaille pourtant centralement sur ce qu'il appelle « le document numérique », montrant que ce dernier modifie nos rapports à l'empirique et propose surtout de nouveaux observables (Rastier 2011). Chez François Rastier et les chercheurs qui travaillent dans sa perspective, l'expression *document numérique* semble cependant rendre compte plus du support que de l'écosystème numérique en ligne ou hors ligne. Je reste donc en attente d'une véritable prise en compte de la matérialité technotextuelle, les analyses de la sémantique des textes et de corpus revenant souvent, comme à leur état naturel, vers des descriptions de la textualité insuffisamment contextualisées dans les milieux natifs. Je m'étends sur le texte ici car le groupe Pédauque en fait le « concurrent » du document, mais il en est de même en analyse du discours et des interactions, et je regrette que pour les disciplines TDI, qui ont pourtant montré depuis les années 1970 la puissance théorique et méthodologique de leurs élaborations, les univers discursifs numériques constituent encore une sorte d'angle mort de la discipline. Il est temps que les linguistes TDI viennent explorer les corpus natifs du web en y installant leurs méthodologies et leurs interrogations, au lieu de venir y prélever par extraction des corpus qui perdent alors leur dimension technolangagière.

BIBLIOGRAPHIE

Achard-Bayle G. & Paveau M.-A., 2012. Réel, contexte, et cognition. Contribution à une histoire de la linguistique cognitive, avec G. Achard-Bayle, *HEL Histoire Épistémologie Langage* XXXIV (1), 97-114.

Adam J.-M., 2006. Autour du concept de texte. Pour un dialogue des disciplines de l'analyse des données textuelles. Conférence plénière d'ouverture des Journées internationales d'Analyse des Données Textuelles (JADT), Besançon, 19-21 avril 2006, en ligne sur *Lexicometrica*, http://lexicometrica.univ-paris3.fr/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf

Aït Saïdi L. & Matthey M., 2012. Pratiques linguistiques de 4 facebookiens kabylophones : un exemple de métissage de langues. Communication au Colloque de la VALS-ASLA : *Le rôle des pratiques langagières dans la constitution des espaces sociaux pluriels d'aujourd'hui : un défi pour la linguistique appliquée,* Université de Lausanne, 1-3 février 2012.

Boch F., Grossmann F. & Rink F., 2007. Conformément à nos attentes... *: les marqueurs de convergence/divergence dans l'article de linguistique*. Revue Française de Linguistique Appliquée, XII-2, 109-122.

Cailleux A., 1953. L'évolution quantitative du langage. *Bulletin de la Société préhistorique de France*, tome 50, N. 9-10, 505-514.

Feyerabend P., 1979 [1975]. *Contre la méthode. Esquisse d'une théorie anarchiste de la connaissance*. Paris : Seuil, coll. Points sciences.

Guilhaumou J., 2002. Le corpus en analyse de discours : perspective historique. *Corpus* 1 : http://corpus.revues.org/index8.html

Habert B.

– 2005a. *Instruments et ressources électroniques pour le français*, Gap/Paris, Ophrys.

– 2005b. Portrait de linguiste(s) à l'instrument. *Texto ! Textes et cultures*, Vol. X-4 : http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html

Guillaud H., 17 décembre 2010. Culturomics : Comprendre les « lois » de la culture. *La feuille*, [blog], http://lafeuille.blog.lemonde.fr/2010/12/17/culturomics-comprendre-les-lois-de-la-culture/, consulté le 10 novembre 2012.

Habert B., Nazarenko A. & Salem A., 1997. *Les linguistiques de corpus*. Paris : A. Colin.

Labbé D.

– 2001. Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*, 8-3, 213-231.

– 2009. *Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan*. Paris : Max Milo.

– site personnel : http://www.pacte.cnrs.fr/spip.php?article54

Labbé C. & Labbé D., 2003. La distance intertextuelle. *Corpus* 2, http://corpus.revues.org/index31.html

Lemercier C. & Zalc C., 2008. *Méthodes quantitatives pour l'historien.* Paris : La Découverte.

Mayaffre D., 2010. *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, mémoire pour l'habilitation, http://tel.archives-ouvertes.fr/tel-00655380

Meillet A. & Cohen M., 1924. *Les langues du monde*. Paris : Champion.

Michel J.-B. et *al.*, 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science magazine*, published online on "Science*express*", http://www.sciencemag.org/content/331/6014/176

Mounier P., 2011. Du discours aux données et retour... *Homo Numericus*, http://www.homo-numericus.net/breve1008, consulté le 3 avril 2012

Mucchielli A. (dir.), 2008. *Dictionnaire des méthodes qualitatives en sciences humaines*. Paris : A. Colin.

Paveau M.-A.

– 2007. Discours et cognition. Les prédiscours entre cadres internes et environnement extérieur. Corela (Cognition, Représentation, langage) : « Contextes, discours, cognitions », http://corela.edel.univ-poitiers.fr/index.php?id=1550

– 2009. Mais où est donc le sens ? Pour une linguistique symétrique. Conférence invitée au deuxième colloque international *Res per nomen*, Reims, 30-31 mai, in Actes prépubliés, p. 21-31.

– 2012a. Activités langagières et technologie discursive. L'exemple du réseau de micro-blogging Twitter. Communication au Colloque de la VALS-ASLA : *Le rôle des pratiques langagières dans la constitution des espaces sociaux pluriels d'aujourd'hui : un défi pour la linguistique appliquée,* Université de Lausanne, 1-3 février 2012, présentation en ligne http://penseedudiscours.hypotheses.org/8338

– 2012b. Les amitiés numériques. Formes technolangagières et polysémie relationnelle. Conférence au Séminaire PHITECO 2012 *Amitié et technologies relationnelles*, Université de Technologie de Compiègne, du 23 au 27 janvier 2012, présentation en ligne : http://penseedudiscours.hypotheses.org/8334

– 2012c. Réalité et discursivité. D'autres dimensions pour la théorie du discours. *Semen* 34, 95-115.

– 2012d. Linguistique et numérique 2. Le désert des blogs francophones. *La pensée du discours* [carnet de recherche], http://penseedudiscours.hypotheses.org/10212, consulté le 10 novembre 2012.

– 2013 (à par.). Genre de discours et technologie discursive. Tweet, twittécriture et twittérature. *Pratiques*, dossier « Modèles et théories du genre en confrontation ».

Pédauque R.T. (collectif), 2003. Document : forme, signe et médium, les re-formulations du numérique. http://archivesic.ccsd.cnrs.fr/sic_00000511

Rastier F., 2011. La mesure et le grain. Sémantique de corpus. Paris : Champion.

NOTES

1. On en trouvera une synthèse et une bibliographie dans Lemercier et Zalc 2008.

2. J'utilise le terme *matière* de préférence à *matérialité* pour des raisons historiques et théoriques : ce que Michel Pêcheux et ses collaborateurs appelaient « matérialités langagières » dans les années 1980 reposait sur une conception du discours qui faisait droit à l'inconscient à travers la notion de « réel de la langue ».

3. J'appelle « disciplines TDI » les disciplines du texte, du discours et de l'interaction, étendues à tous les travaux qui prennent pour objets des énoncés empiriques, c'est-à-dire non fabriqués dans une perspective expérimentale : sociolinguistique, anthropologie linguistique, etc.

4. Je parle de *technolangagier* quand les signes linguistiques ne sont plus purement langagiers mais assemblés ou métissés avec du technique, de manière non distinguable.

5. Dans cet article, j'utilise la notion d'intrumentation principalement sous ses aspects quantitativistes, et j'écarte de l'analyse ses usages descriptifs ou opératoires qui n'ont pas nécessairement pour but d'établir des comptages. C'est une réduction certaine de la notion, que j'assume comme telle.

6. Je travaille moi-même « à la main » et « à l'humain » et je ne fournis pas de chiffres pour soutenir ce que j'avance, ce que je devrais faire. Je n'ai pas réalisé d'enquêtes pour chiffrer la

proportion des travaux qualitatifs et quantitatifs en sciences du langage. Il serait sans doute possible de réaliser des sondages sur les thèses présentées au CNU chaque année pour l'obtention de la qualification, ou sur l'ensemble des articles publiés sur une année dans les revues de linguistique francophones, par exemple.

7. Le non-humain englobant les objets naturels et artificiels, et les animaux.

8. Un document sur le web est une collection de pages individuelle liées entre elles par des hyperliens, dotées d'espaces optionnels de défilement, de texte, et éventuellement d'images et d'autres médias intégrés.

9. On appelle *redocumentation* sur le web l'opération qui consiste à rassembler et ordonner des données de manière à en faire un ensemble lisible de documents : les procédures de visualisation de données, de journalisme de données ou de curation de données sont des procédures de redocumentation.

RÉSUMÉS

On trouve peu d¹interrogations historiques et épistémologiques sur la question des approches qualitatives et quantitatives en analyse des textes et des discours. Dans cet article, je propose de questionner la symétrie entre qualitativisme et quantitativisme, la nature du savoir linguistique produit, la prise en compte du contexte et la pertinence de l'une ou l'autreapproche sur les corpus de discours natifs en ligne.

There is not much historical and epistemological works on the issue of qualitative and quantitative approaches in the field of text and discourse analysis. In this paper, I propose to consider the symmetry between qualitativism and quantitativism, the nature of the linguistic knowledge produced, the way the context is taken into account and the relevance of either approach regarding web native discourse corpora.

INDEX

Mots-clés : discours natifs en ligne, épistémologie, linguistique symétrique, numérique, technolangage
Keywords : digital, epistemology, symmetrical linguistics, technolanguage, web native discourse

AUTEUR

MARIE-ANNE PAVEAU

Université Paris 13 – Sorbonne Paris CitéCENEL