Introduction.

Pour une science des textes instrumentée

Mathieu Valette
ATILF (CNRS, Nancy)
mvalette@atilf.fr

La *linguistique de corpus* ne sera, selon toute vraisemblance, jamais établie en discipline académique. Aujourd'hui, nombre de linguistes, quels que soient leur discipline ou leurs objets d'étude, sont conduits à constituer des corpus numériques et à les étudier au moyen d'outils logiciels chaque année plus nombreux, sophistiqués et conviviaux. La banalisation de l'outil désenclave ainsi des pratiques longtemps réservées à une petite minorité que l'informatique ne rebutait pas.
Mais cette évolution technologique, si elle peut avoir une incidence méthodologique (par exemple et minimalement, en substituant aux exemples construits des exemples attestés), n'a pas pour autant un impact fort sur les théories ni sur la définition des objets de la linguistique : à la morphologie, les corpus de mots ; à la syntaxe, les corpus de phrases ; aux théories énonciatives, les corpus d'énoncés. Et bien que tous ces objets d'étude proviennent de textes, ceux-ci ne sont que rarement considérés comme objet de science dans ces contextes disciplinaires. Ils sont réduits, par défaut, au statut préscientifique de ressource – un matériau brut dont la qualité est déterminée par la seule présence, après raffinage, de l'objet étudié. On collecte ainsi de l'indénombrable : *du* texte ou *du* corpus.

De la linguistique de corpus à une science des textes instrumentée

Or, le texte fait l'objet, avec cette fameuse société de l'information, d'un intérêt nouveau. Sa problématique s'articule en effet avec celle, récente, du document numérique, lequel est, pour beaucoup, le vecteur d'une révolution aussi importante que jadis le passage du *volumen* au *codex*. C'est peu dire que l'accroissement des données textuelles numérisées est actuellement soutenu, du fait d'Internet évidemment, mais aussi de la Gestion Electronique de Documents (GED). Ces nouveaux modes de

---

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

production, de stockage et d'accès au document génèrent, outre des dépenses énergétiques considérables, de nouvelles questions et de nouvelles problématiques en termes d'analyse et d'indexation des contenus, de recherche d'information et d'interprétation assistée.

Les linguistiques du texte, jusque-là souvent cantonnées à l'analyse des textes littéraires ou politiques aux genres globalement bien décrits par la tradition, se trouvent confrontées à une grande variété de discours et de genres nouveaux, indéterminés, polymorphes et en permanente évolution[1] qu'il leur appartient de caractériser. Que ces discours et ces genres soient traces de nouvelles pratiques sociales ou modernisation de pratiques anciennes, il apparaît crucial pour la linguistique, science humaine et sociale, de prendre position face aux enjeux théoriques et méthodologiques naissants, et de ne pas laisser à d'autres disciplines (sciences de l'information et de la communication, ingénierie des connaissances, etc.) le soin de décrire, seules, ces nouveaux objets sémiotiques.

Parmi les linguistiques du texte, les propositions théoriques de F. Rastier (sémantique interprétative, sémantique textuelle) participent activement à ce débat (Rastier 1987, 2001). Ayant pour objet empirique le texte et non le mot, la phrase ou l'énoncé, traditionnellement privilégiés, cette linguistique-science des textes renoue avec une tradition rhétorique et herméneutique oubliée du XXème siècle et se concentre sur l'étude de la textualité, des genres textuels, des discours et de leurs corollaires (cohésion textuelle, intertextualité, etc.). Son appareil théorique est depuis le début des années 90 éprouvé par la linguistique de corpus[2], le TAL[3] et plus récemment par la Recherche d'Information[4].
L'association des outils logiciels aux outils théoriques et conceptuels peut avoir différentes conséquences : (i) la validation logico-mathématique (établissement d'un « modèle » informatisé de la théorie) – valorisée dans

---

1 Par exemple, dans le domaine de l'auto-édition, les *pages perso*, florissantes il y a quelques années, tendent aujourd'hui à se marginaliser tandis que les blogs, réputés interactifs, se développent fortement. Ils associent généralement les billets (ou notes, ou *posts*) d'un internaute ou d'une communauté d'internautes-auteurs et quelques commentaires d'internautes-lecteurs).

2 Lire (Bourion 2001), (Loiseau 2006), (Poudat 2006).

3 Lire (Beust 1998), (Thlivitis 1998) (Bommier-Pincemin 1999), (Perlerin 2004), (Rossignol 2005).

4 Lire (Valette 2004), (Mauceri 2007), (Valette & Slodzian 2008).

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

certaines traditions linguistiques comme la syntaxe générative ou la linguistique cognitive, ce mode de validation, né avec l'informatique dans le sillon cybernétique, relève d'une esthétique scientifique souvent mal adaptée aux sciences humaines et sociales ; (ii) la validation pratique (par des applications logicielles). Elle présente l'intérêt de confronter les scientifiques aux demandes sociales mais elle demeure assujettie aux passions technoscientistes contemporaines ; enfin (iii) le déploiement de l'objet de recherche : l'instrumentation, constitutive de la linguistique de corpus, donne lieu à ce que nous pourrions appeler son « cercle vertueux ». Les grandes masses de données textuelles ou documentaires nécessitent, pour être analysées et décrites, des dispositifs expérimentaux et des instruments *ad hoc*. Cette instrumentation permet de construire de nouveaux observables qui seraient demeurés invisibles autrement.

Propositions

C'est autour de ce concept d'observable et des vertus afférentes de l'instrumentation que s'articule cette livraison de la revue. Nous entendons faire le point sur certains des derniers développements de la linguistique des textes lorsque celle-ci a recours à des instruments de mesure. L'opus se focalise sur différents aspects porteurs pour la linguistique tant d'un point de vue théorique et épistémologique que dans la perspective de son applicabilité à des besoins sociaux, culturels et économiques aujourd'hui bien identifiés.

F. Rastier inaugure ce recueil par une réflexion sur l'articulation entre les concepts de texte, de document numérique, de donnée et de métadonnée. Initiant une discussion entre la linguistique des textes et l'ingénierie des connaissances, il oppose les connaissances du Web sémantique, fondées sur une approche ontologique *a priori* des concepts et ignorante des pratiques sociales qui ont permis la production des documents pourtant porteurs des concepts, et les *connaissances sémiotiques* (incluant le textuel) soumises aux évaluations et aux validations induites par la pratique. Loin des sentiers battus, il plaide contre le Web sémantique et pour une sémantique du Web restituant les contextes de production.

Les trois contributions suivantes exposent des réflexions nourries sur les relations entre interprétations et textualité. A cette fin, les auteurs convoquent et illustrent chacun à leur manière la notion d'*isotopie*.

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

I. Kanellos et Chr. Mauceri enquêtent sur les possibilités de réalisation d'une plateforme d'analyse interprétative des données conforme aux propositions de l'herméneutique traditionnelle. Leur projet est « de construire un outil en vue de donner corps, par son maniement, au cercle herméneutique de l'interprétation des données ». Ils combinent à cette fin des outils interprétatifs tels que l'isotopie et des méthodes statistiques éprouvées comme l'analyse sémantique latente. L'un de leurs ambitieux objectifs est la réintroduction d'une autorité sémantique qui coopére avec le calcul et de lutter ainsi contre l'« attitude de prédation » des modèles et des techniques de calcul sur l'objectivité scientifique.

Les contributions de D. Mayaffre et S. Loiseau approfondissent le concept de contextualité. Le premier soutient notamment l'hypothèse que la cooccurrence en est la forme minimale. Tout en en détaillant opportunément l'histoire et les usages au sein de l'analyse des données textuelles, dans le domaine anglo-saxon et dans la tradition française, il discute de l'incidence du passage d'une statistique occurrentielle à une statistique cooccurrentielle. Selon lui, une cooccurrence, si elle est observée, est déjà sémantique : la cooccurrence peut en effet être perçue comme la première forme de contextualisation d'un mot par les autres. L'enjeu, pour D. Mayaffre, est de reconstituer la « trame lexicale » complexe ou les « entrelacements » lexicaux sous-jacents dans le corpus. Il ambitionne ainsi de contrôler la recherche des isotopies, des réseaux sémantiques ou des thèmes sans avoir recours à la seule intuition.

S. Loiseau développe des propositions voisines mais en déplace sensiblement les arguments. Selon lui, les interactions entre les différents types de normes et les interprétations liées à plusieurs niveaux d'analyse du texte constituent deux formes de contextualité à approfondir. S. Loiseau s'intéresse donc aux corpus multi-annotés qui articulent plusieurs niveaux de descriptions (morphologique, lexical, morphosyntaxique, syntaxique). Pour Loiseau, les observables issus d'annotations multiples permettent de décrire des normes linguistiques comme les discours et, de la sorte, d'accéder à la complexité empirique du palier de la textualité.

La contribution de B. Pincemin et ses collaborateurs est singulière : il s'agit de faire une synthèse des pratiques textométriques de quelques linguistes travaillant sur des corpus de textes médiévaux en analysant à la fois

---

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

l'historique des requêtes soumises au logiciel d'interrogation et l'archive des questions des utilisateurs. Les auteurs en dégagent une interprétation linguistique des stratégies d'interrogation développées dans le cadre d'une instrumentation comportant des aspects classiques (moteur de recherche) et des aspects plus élaborés (calculs statistiques). Ce travail que l'on qualifierait ailleurs d'« analyse métier » a des incidences dans la formation, la documentation des outils et dans l'ergonomie logicielle.

Il eût été regrettable de plaider pour une science des textes sans prendre en compte ce qui constitue historiquement un de ses fondements : la diversité des langues. Le projet de la linguistique oscille entre la description des variations inter-langues et l'observation d'universaux ou, plus récemment, d'invariants. M. Slodzian observe que les nouvelles problématiques du document numérique et des réseaux électroniques sont confrontées à cette question. Le multilinguisme y est vu tantôt comme un obstacle, tantôt comme un atout. Selon elle, c'est la diversité linguistique et culturelle comme phénomène sémiotique fondamental qui est en jeu. Suivant l'orientation choisie, les programmes linguistiques sont en effet diamétralement opposés : ceux qui, au nom de l'efficacité, prônent la « débabélisation du monde » souhaitent le développement d'un instrument de communication, autrement dit un *« interlinguisme »* réducteur. Ceux qui voient dans la variété et la différence la condition même de la vie culturelle des sociétés auront pour objectif le *« translinguisme »*.

Par endroits, l'article de M. Slodzian fait écho à celui de F. Rastier car ils traitent tous deux, à leur manière, des pressions multiples exercées sur les langues par la fameuse mondialisation économique, politique et culturelle (pesée d'un très petit nombre de langues véhiculaires, normalisation et appauvrissement des échanges linguistiques, dévaluation des langues peu parlées, etc.) et de leurs conséquences pour une linguistique, *science impliquée*.

Enfin, J.-M. Daube apporte des éléments de réponse empiriques aux questionnements de M. Slodzian. Il expose une recherche en lexicologie textuelle visant la réalisation de lexiques bi- et trilingues. Il s'agit d'identifier et de recenser, dans une perspective lexicographique, des lexies à partir de corpus homogènes typés par domaines. Constatant les limitations des corpus parallèles alignés (corpus de textes traduits), J.-M. Daube discute de l'opposition *corpus parallèle vs. corpus comparable*. Il

---

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

observe que la problématique de la constitution et de l'exploitation des corpus comparables n'est, à l'heure actuelle, qu'esquissée.

Ces sept contributions tracent à grands traits un parcours au sein de ce qui constitue une science des textes moderne où les formes documentaires sont considérées non pas seulement comme un nouveau matériau succédant au précédent mais comme les vecteurs de nouvelles pratiques. Loin d'une technophilie ravie, l'établissement d'une linguistique-science des textes instrumentée pose de façon critique – et pratique – sa relation à l'outil, son rapport à la Technique, et aux dangers d'une substitution non réflexive voire, osons l'épithète, totalitaire.

Références bibliographiques

Beust, P. (1998) *Contribution a un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines*, Thèse de doctorat, Caen.

Bommier-Pincemin, B. (1999) *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Paris IV Sorbonne.

Bourion, E. (2001) *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Nancy 2.

Loiseau, S. (2006) *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*, Thèse de doctorat, Paris X-Nanterre.

Mauceri, Chr. (2007) *Indexation et isotopie : vers une analyse interprétative des données textuelles*, Thèse de doctorat, ENSTB.

Perlerin, V. (2004) *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat, Caen.

Poudat, C. (2006) *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse de doctorat, Orléans.

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.

Rastier, F. (1987) *Sémantique interprétative*, Paris, PUF.

Rastier, F. (2001) *Arts et sciences du texte*, Paris, PUF.

Rossignol, M. (2005) *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat, Rennes 1, disponible sur http://www.texto-revue.net.

Thlivitis, T. (1998) *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de doctorat, Rennes 1.

Valette, M. (2004) « Sémantique interprétative appliquée à la détection *automatique de documents racistes et xénophobes sur Internet », Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, P. Enjalbert et M. Gaio, eds, Entropia, Paris, 215-230.

Valette, M., Slodzian, M. (2008) « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*, A. Condamines & Th. Poibeau, éds., *Revue Française de Linguistique Appliquée* (volume XIII-1 / juin 2008), 119-133.

Fac-similé de l'introduction de M. Valette, éd., (2008) *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, Syntaxe & Sémanique, n°9, 2008, 9-14.