

L'ACCÈS SÉMANTIQUE AUX BANQUES TEXTUELLES

François RASTIER

Institut National de la Langue Française
Centre National de la Recherche Scientifique
Grille d'honneur, Le Parc, F. 92211 Saint-Cloud
courriel : lpe2@ext.jussieu.fr / toile : //www.msh-paris.fr/texto/

Abstract : Semantic access to textual databases demands further research on its theoretical foundations.Using a semantics of norms we must approach the diversity oftexts through the diversity of genres and the social practices. Coding of genres is a precondition for improved textbase ergonomy. Typological criteria are intralinguistic (structures and units) as well as extralinguistic (text purposes and situations).Text semantics has defined its units and relations independently of the compositionality principle : themes and topoi, motifs and dialectical functions, etc. Computerized access to large corpora offers new ways to study textual stereotypes. Assisted interpretation involves not only creating enriched sub-corpora but also defining the system of pertinent parameters which varies with each application. The development of non-linear readings has renewed the interpretative processes that are peculiar to traditional uses of written texts.

Key-Words : Semantics, Corpora, Interpretation, Philology, Textual Databases.

1. SÉMANTIQUE DE CORPUS

Le recueil et l'étiquetage de grands corpus informatisés permet à la linguistique de définir un nouvel objectif : l'accès sémantique aux banques textuelles. Ce type d'application fait l'objet d'une large demande sociale mais exige un approfondissement théorique en sémantique interprétative. D'une part, elle doit renouer avec la philologie, qui reste à la base de tout traitement des textes ; d'autre part, étudier les pratiques interprétatives (en exploitant

ISBN: 0 08 043 438X

l'herméneutique philologique) pour proposer des formes de codage et les exploiter. Tout codage résulte en effet d'une interprétation, et guide ou contraint les interprétations futures.

Nous présenterons ici l'orientation générale de travaux menés au sein de l'équipe Sémantique des textes (Inalf-Cnrs).

2. TYPOLOGIE

Comme tout texte procède d'un genre, et tout genre d'un discours, il convient de rapporter, par une sémantique des normes — et non plus seulement de la "langue" —, la diversité des textes à la diversité des genres et des pratiques sociales. Si la description linguistique traditionnelle lisse les genres pour créer l'illusion d'une langue générale et neutre, le codage préalable des genres reste crucial pour restituer la diversité des normes et des usages, et assurer ainsi l'ergonomie de l'accès banques textuelles (cf. le projet Silfide). Les critères proposés par la *Text Encoding Initiative*, qu'ils soient fonctionnels (*plaire, informer, exprimer, persuader*) ou référentiels (*"factualité" ou fictionnalité*) sont trop généraux, sans fondement linguistique, et ne correspondent ni à des discours ni à des genres. Aussi proposons-nous des critères de typologie intralinguistiques (structures et unités) et extralinguistiques (objectifs et situation des textes) : dans les deux cas, la contextualisation opérée par la sélection du corpus conditionne les résultats d'analyse. Cela permet d'utiliser des méthodes contrastives et d'étudier ainsi les normes sémantiques à l'œuvre dans les textes.

3. STRUCTURES ET UNITÉS

Faute de compositionnalité du sens, la problématique logico-grammaticale s'applique mal aux textes : les procédures de segmentation utilisant des balises sont utiles pour traiter de l'expression, mais sans plus. Aussi, la sémantique des textes a dû (re)définir d'autres formes d'unités et de relations qui en sont indépendantes : isotopies, thèmes et topoï, motifs et fonctions dialectiques, etc. (cf. l'auteur, 1989, 1995). Les isotopies sont des fonds sémantiques, les thèmes et topoi des formes (décrites comme des molécules sémiques, petits réseaux sémantiques dont les nœuds sont des sèmes et les liens des cas). Des rapports forme/fond du même ordre peuvent être décrits dans les autres composantes sémantiques dialectique (narrative) ou dialogique (modale), voire tactique (séquentielle). Nos hypothèses théoriques de base postulent ceci :

(i) Sans égard pour le dualisme, l'*unité contenu/expression* est établie par les parcours interprétatifs : par exemple, dans un corpus romanesque, E. Bourion a ainsi pu confirmer la corrélation entre des noms de sentiments et les ponctuations dans les contextes ou ces noms apparaissent.

(ii) La *diffusion sémantique* qui rend compte des phémomènes d'isotopie. Comme tout trait sémantique est actualisé et *a fortiori* propagé à partir et en fonction du contexte immédiat et lointain, les contextes manifestent des redondances locales multiples.

De là découlent des propositions méthodologiques :

(i) Les cooccurrents d'un mot-pôle sélectionnés par le test statistique de l'écart réduit peuvent être qualifiés sémantiquement et devenir des corrélats sémantiques, c'est-à-dire des sémies voisines comportant au moins un trait sémantique commun. Les unités de rang supérieur, comme un thème ou un acteur sont alors caractérisés par des cliques de corrélats. On peut ainsi passer, non sans conditions, du quantitatif au qualitatif et du lexical au textuel.

ISBN: 0 08 043 438X

(ii) Comme le paragraphe ou du moins la période sont les unités de base de la textualité (l'alinéa inhibe les propagations), l'interrogation par mots doit être abandonnée au profit de l'interrogation texte-texte : ainsi, par exemple le système DECID de X. Lemesle et B.Bommier-Pincemin (EDF/DER) systématise et applique la technique herméneutique des passages parallèles.

(iii) La recherche documentaire doit être conçue une exploration des textes pour l'aide à l'interprétation, par sélection de sous-corpus à pertinence enrichie, en fonction de la tâche en cours.

4. STÉRÉOTYPIE ET DOXA

L'accès à de grands corpus permet d'étudier avec des moyens nouveaux la stéréotypie textuelle et les normes de la doxa. L'exemple le plus simple est celui de la canonicité : dans le corpus roman 1830-1970 de la banque Frantext, qui compte environ 350 œuvres, on trouve seulement 5 sortes de fractions de seconde, et 12 nombres de secondes (sur une infinité théoriquement possible). Sur 4488 mentions d'âge — 2650 hommes (59 %) et 1838 (41%) femmes — , certains âges n'apparaissent pas : 41 ans pour les femmes, 49 ans pour les hommes, 71 ans ou encore 92 ans ; d'autres sont sur-représentés, par exemple 15, 18 et 20 ans pour les deux sexes ; 16 ans pour les personnages féminins (résultats dûs à N. Deza). Dans le roman français, on a presque toujours vingt ans.

Par ailleurs, l'étude de la stéréotypie permet de lier les occurrences de lexies à des formes textuelles : par exemple, dans le même corpus, *au pied de* (singulier) est toujours un localisant dans une description, *aux pieds de* (pluriel) appartient toujours à un récit d'imploration ou de vénération (résultats dûs à E. Bourion). Dans un lexique de ce corpus, cette lexie devrait donc figurer sous deux entrées différentes.

On peut considérer que la concrétisation la plus simple d'une doxa (ou système axiologique) est un lexique : la doxa commande en effet la constitution des classes lexicales minimales (taxèmes), et par là la définition différentielle des sémèmes et des sèmes en leur sein. La méthodologie de construction de lexiques ouvre un domaine d'application crucial pour les traitements automatiques du langage (cf. Cavazza, 1997).

5. PARCOURS INTERPRÉTATIFS

Le développement des lectures non-linéaires est en voie de renouveler voire de bouleverser les parcours interprétatifs propres aux usages traditionnels de l'écrit. Au delà de la problématique logico-grammaticale, les systèmes d'assistance à l'interprétation doivent pour décrire les formes textuelles s'appuyer sur la problématique rhétorique / herméneutique. En effet :

(i) Les objectifs et contraintes pratiques, différenciés en discours, genres et styles, configurent les formes textuelles.

(ii) Elles sont caractérisées par des inégalités qualitatives (masquées par les théories propositionnelles) codifiées ou non. Parmi ces inégalités, il faut noter les degrés de concentration des formes selon que leur manifestation est compacte ou diffuse.

(iii) Un autre régime d'inégalités qualitatives, la pertinence par rapport à la tâche en cours réorganise ces saillances relatives des formes sémantiques.

ISBN: 0 08 043 438X

Les systèmes d'assistance à l'interprétation procèdent par extension (recontextualisation) ou restriction (caractérisation). Leurs fonctionnalités sont globalisantes quand ils permettent des mouvements d'extension ou de restriction du corpus de référence ou du corpus de travail. Fournir une assistance à l'interprétation conduit ainsi à créer des sous-corpus enrichis, et d'autre part à définir des régimes de pertinence variables selon les applications.

Elles sont localisantes quand elle permettent des discrétisations (soit par analogie : recherche des passages parallèles, comme dans l'interrogation texte/texte, soit par contraste). L'identification de formes textuelles, qui relève de la reconnaissance de formes et non du calcul, peut alors se faire par sommation qualitative.

Les parcours interprétatifs élémentaires qui décrivent les opérations d'actualisation (et de virtualisation) de sèmes sont préactivés par des parcours globaux. Le paramétrage de ces rapports global / local reste un problème ouvert : il diffère vraisemblablement selon les genres et les discours.

RÉFÉRENCES

Berkenkotter, C. & Huckin, T. N. (éd.) (1995). *Genre Knowledge in Disciplinary Communication*. Hillsdale (N. J.) Lawrence Erlbaum.

Bhatia, V. K. (1993). *Analysing Genre : Language Use in Professionnal Settings*. Londres, Longman.

Biber, D. (1988). *Variations across Speech and Writing*, Cambridge, CUP.

Cavazza, M. (1997). Sémiotique textuelle et contenu linguistique, *Intellectica*, 23, pp. 53-77.

Church, K. and Hanks, P. (1989). Introduction to the special issue on computational linguistics using large corpora, *Computational linguistics*, 19, 1, pp. 1-24.

Jacobs, P. (1990). To Parse or not to Parse: Relation-Driven Text Skimming, *Proceedings of COLING'90 Conference*, Helsinki.

Lardet, P. (1992). Travail du texte et savoirs des langues, in Auroux, S. (éd.) *Histoire des idées linguistiques*. Bruxelles, Mardaga, t. II, pp. 187-205.

Mettinger, A. (1994). *Aspects of Semantic Opposition in English*. Oxford, Clarendon Press.

Rastier, F. (à paraître) Herméneutique matérielle et sémantique des textes, in Salanskis, J.-M. et al. (éds) *Herméneutique : textes, sciences.*Paris, PUF.

Rastier, F. (1987). *Sémantique Interprétative*. Paris, Presses Universitaires de France.

Rastier, F. (1989). *Sens et Textualité*. Paris, Hachette.

Rastier, F. (1991). *Sémantique et Recherches Cognitives*. Paris, Presses Universitaires de France.

Rastier, F. (1995). *L'analyse thématique des données textuelles — L'exemple des sentiments*. Paris, Didier.

Rastier, F.(1997). Problématiques du signe et du texte, *Intellectica*, 23, pp. 7-53.

Rastier, F., Cavazza, M. et Abeillé, A. (1994). *Sémantique pour l'analyse*. Paris, Masson.

Swales, J. M. (1990). *Genre Analysis. — English in Academic and Research Settings*. Cambridge, Cambridge University Press.

ISBN: 0 08 043 438X