Pr. Chiali FZ
Université d'Oran

14. Traitement informatique du texte littéraire

Dans ce travail, nous tentons d'expliciter le traitement informatique du texte littéraire et les procédés qui œuvrent à ses mises en co(n)textes dont les visées seraient l'interprétation.

Plan

1. Traitement informatique du texte littéraire

Avant d'expliciter les aspects méthodologiques de notre recherche, il importe de préciser le rôle que nous voudrions assigner aux traitements informatisés pour l'étude sémio-linguistique d'un texte littéraire. Il reste vrai que la logique binaire de 1 et de 0, apparaît de prime abord comme un outil inadapté et insuffisant pour partir à la recherche d'une signification fluente, sans cesse réévaluée. En fait, plusieurs reproches peuvent être adressées à la pratique informatisée du texte littéraire. Tout d'abord, même si l'on peut produire à l'aide de l'outil informatique, des données quantitatives de tous ordres, on ne saurait traduire informatiquement certaines approches sémiotiques qu'au prix de programmes très complexes. Le risque serait dans ce cas que l'outil n'altère le concept qu'il est censé mettre en pratique. Ainsi, il est primordial de veiller à ce que le concept prime sur la méthode et l'outil, ce qui implique des mises au point tout au long de l'analyse.

Une pratique informatisée ne garantit pas à elle seule la scientificité de la démarche. C'est alors que les pratiques lexicologiques ont posé le "doute" comme un postulat de base. Celui-ci impose au chercheur de vérifier des informations pour pouvoir les valider. La finalité de la démarche n'est pas de saisir la *vérité* d'un texte littéraire mais de rendre compte de sa réalité objective. En fait, il s'agit à partir de diverses corrélations ou convergences, de formuler des hypothèses qu'il faudra ensuite valider ou infirmer en réunissant un faisceau d'indices significatifs dont la pertinence est un problème de taille. Pierre Lafon, spécialiste des traitements lexicométriques, insiste lui-même sur les dangers d'une pratique statistique assistée ou non par l'outil informatique.

Quel que soit le modèle utilisé, nous sommes toujours exposés à deux types de dangers : certains faits statistiquement significatifs n'ont aucune interprétation linguistiquement pertinente ; d'autres faits jugés a priori pertinents peuvent échapper à la statistique. Les résultats statistiques ne sont donc pas toujours à prendre pour argent comptant. [35]

Ces deux types de dangers nécessitent prudence et modestie : l'informatique est un outil qui ne prétend pas tout constater et *a fortiori* tout expliquer. Par conséquent il faut que les indices obtenus, soient statistiquement pertinents et linguistiquement interprétables, pour que progresse l'analyse sémiotique du texte. Cette

[35] Pierre Lafon, *Mots* N°2, Paris, C. N. R. S, 1981, p. 182.

problématique a ouvert un long débat où l'on accusait à tort les pratiques statistiques (enquêtes lexicométriques, et autres) soit de mettre en évidence des signes patents, que le lecteur perçoit d'emblée ; soit au contraire de souligner des faits non-perceptibles. En fait, les procédures de la pratique statistique tentent de montrer le mécanisme des structures, notamment répétitives, inconsciemment perçues, ainsi que le rôle qu'elles jouent dans la construction du sens.

Une autre difficulté, extérieure à la pratique statistique informatisée, est l'interprétation des résultats. Il est vrai que cette phase demeure inévitablement subjective, incertaine mais nécessaire. Pour en limiter les risques de dérapage, les contresens, il est judicieux de multiplier les analyses et de confronter les résultats. Le traitement informatisé peut être adapté au texte littéraire. Il se révèle puissant et capable de produire un examen exhaustif des corpus, car l'informatique impose la rigueur dans les concepts et les méthodes, rigueur aussi dans la restitution des données : *la certitude d'une lecture cohérente et constante en tout point du texte, régie non par l'intuition mais étalonnée par une analyse quantitative et intégrale du corpus*. De plus s'agissant d'informatique : *la mécanique mise en œuvre contraint à la constitution d'objets précisément défini*[36].
Les pratiques statistiques n'offrent pas de vérité mais révèlent au chercheur une série d'hypothèses, de pistes à entreprendre. Si elles se présentent comme des outils fructueux, c'est plus par les questions qu'elles posent que par les réponses apportées : *Le but de la statistique est de permettre de déceler dans une masse d'informations des faits qui seront la source de réflexion.* [37]

C'est dans cet esprit que nous envisageons les pratiques statistiques dont le traitement informatisé est une sorte de préparation du corpus : traitement logiciel, analyse des données, hypothèses interprétatives puis retour au texte, nouvelles questions, nouveaux outils, adaptation de nouvelles méthodes. La difficulté reste de savoir ce qu'on cherche et dans quel but car :
La méthode d'exploitation est éminemment ouverte. Elle n'impose pas de conclusions mécaniques, déterministes, toutes faites. Les chiffres fournis par l'ordinateur ne constituent pas une fin en soi mais nous renvoient avec insistance au texte. Leur fonction est de provoquer l'intervention souveraine du sujet humain qui relaie alors et dépasse la machine, parce que seul, il possède la compétence linguistique et stylistique. Par sa grande valeur heuristique, par les hypothèses suggérées, ils favorisent une nouvelle lecture, nourrissent un contact enrichi avec le corpus, sans entraver l'intuition ni restreindre la liberté bien comprise du chercheur face au texte où le poète mot à mot a vaincu le hasard. [38]

2. L'outil en mode d'interprétation

L'on sait que l'outil configure l'objet : ainsi, l'utilisation de l'outil automatique induit un certain nombre de préconstruits d'analyse qui reformulent la question de la co(n)textualisation, en la posant non plus en termes de cadre interprétatif *a posteriori*, mais en termes de conditions interprétatives *a priori*.
Les opérations interprétatives préalables ciblent des objets d'analyse prédéterminés et souvent constitués en listes closes. Car le balisage des corpus, tout comme le choix des formes linguistiques dont on observe les cooccurrences, relèvent déjà d'une

[36] Claude Condé, « Variations et Mises en formes. » *In Semen* N°7, Université de Besançon, 1992, p. 7/11.
[37] Jean-Philippe Massonie, *Analyse informatisée des textes*, Université de Besançon, 1990, p. 14.
[38] Michel Juillard, « Etudes quantitatives des champs sémantiques et morpho-sémantiques dans une œuvre littéraire*. » In la recherche française par ordinateur*, Genève-Paris, Slatkine-Champion, 1985, p. 240.

opération interprétative, comme le montrent notamment les difficultés de repérage de certaines formes de discours rapporté : c'est le cas pour certaines formes marquées mais dont l'interprétation n'est pas univoque (guillemets par exemple).. Pour les formes non marquées, il faut faire appel à une interprétation préalable, c'est-à-dire à une analyse qualitative.
Le balisage des formes syntaxiques, la lemmatisation, la racinisation et le choix des formes lexicales à soumettre à l'analyse textométrique préconfigurent les données. Or, non seulement un nombre considérable de formes pose des problèmes de repérage, mais de plus cette manière d'approcher un texte ou un corpus ne prend pas en compte le fonctionnement en réseau des formes linguistiques, notamment lorsqu'il s'agit d'étudier la cohérence / cohésion des textes. Car, si l'identification quantitative de certaines formes linguistiques est pleinement justifiée pour l'étude de ces mêmes formes, elle ne l'est plus pour l'analyse du texte comme unité linguistique.

3. Contexte et Cotexte

Si les formes linguistiques peuvent mettre en place des éléments de leur propre cadre interprétatif, en configurant l'interdiscours – dans cette optique, on dira avec Guilhaumou (2002 : 22) que le discours contient ses propres ressources interprétatives –, leur interprétation en sus peut nécessiter une co(n)textualisation.
On définira le cotexte comme l'environnement textuel immédiat d'une forme linguistique à interpréter. Le cotexte pertinent pour l'analyse peut prendre des dimensions variables ; il est constitué d'une série d'emboîtements ayant pour frontière le texte. Ainsi, dans certains cas, on élargira le cotexte jusqu'au paragraphe ou au type de rubrique, voire au paratexte.
En même temps, si on définit le texte comme une unité linguistique complexe caractérisée par la cohérence et la cohésion interne, c'est la question du contexte qui se pose. Par exemple, Charolles (1995) insiste sur la pertinence contextuelle en tant que condition et déclencheur de la cohérence d'un texte. De même, plus récemment, Cornish (2006, en ligne) souligne-t-il l'importance du contexte pour transformer un texte en discours : « De toute manière, le texte est souvent, sinon toujours, à la fois incomplet et indéterminé par rapport au discours qui peut en être dérivé à l'aide d'un contexte ».

4. Le corpus : un objet signifiant construit

Le corpus contribue directement à la construction de l'objet linguistique. De fait, ce que l'on observe, ce sont des ajustements successifs entre l'objet et le lieu de son observation.
On insistera aussi sur le fait que l'analyse automatique, de par sa vision en surplomb, contribue à transformer le corpus en contexte interprétatif, grâce aux régularités génériques qu'elle met en évidence et qu'il devient ensuite possible d'appliquer à l'analyse d'un texte afin d'en évaluer la cohérence et de dégager les principes sur lesquels cette dernière prend appui.
Cependant, si le corpus peut se constituer en outil d'interprétation d'un texte, il n'épuise pas pour autant le contexte, qui reste une donnée ouverte. De même, le fait que le discours construise son propre hors-discours à travers une série de formes linguistiques garantit un garde-fou assez fiable contre une contextualisation aléatoire du corpus.
C'est-à-dire qui intègre la dimension verticale de l'interdiscours.
Pour le dire en d'autres termes, l'objectif est de ne pas « manquer le texte » : « Manquer le texte ».
En effet, le texte cotextualise les formes linguistiques en étant lui-même co(n)textualisé par les genres. Deuxièmement – et en lien direct avec le premier

point –, on distingue plusieurs niveaux de co(n)textualisation qui sont appelés à faire le lien entre texte, corpora et genre. Le corpus étire et dilue partiellement les frontières du texte en proposant de l'identifier non plus par rapport à ses intérieurs (cohésion, paragraphes, etc.) mais par rapport à ses extérieurs. On a orienté le texte vers ses extérieurs, ce qui permet de le préserver en tant qu'unité d'analyse sans le « noyer » dans le corpus. L'analyse automatique permet de configurer ces liens allant du corpus au texte.

5. Explorations informatiques : ADT ou ATO

Depuis une bonne dizaine d'années dans le champ francophone, une bifurcation s'est opérée entre les environnements qui se veulent d'*analyse de données textuelles* (ADT) ou *analyse de textes par ordinateur* (ATO) pour reprendre une terminologie québécoise. Le choix se porte sur le statut donné aux résultats des analyses, toutes plus ou moins liées aux méthodes statistiques multidimensionnelles, le plus souvent d'essence benzécriste, et ce statut lui-même renvoie au type d'application visée.

L'essentiel n'est pas de décorer des résultats « objectifs », obtenus par l'application d'algorithmes impassibles à des données réputées les plus fiables et représentatives possible, par des citations pertinentes (les « phrases significatives », par exemple, qui sont recrutées statistiquement pour être censées concentrer au mieux les propriétés de classes calculées). Si le *retour au texte* se limite à cela en substance, il ne mérite pas encore pleinement son nom.

L'essentiel me semble au contraire de configurer un nouveau mode de lecture du plein texte en employant les extractions que permet la statistique probabiliste, comme des outils d'orientation. Ce que les sciences des textes peuvent et doivent apporter à l'Analyse de Discours (c'est-à-dire ce qu'elles doivent s'apporter mutuellement, travailler ensemble) c'est justement ce mixte d'ambition heuristique et de prudence herméneutique.

Conclusion

Il ne faut pas perdre de vue le fait que, aussi bien le corpus que l'objet d'analyse, ainsi que les différents niveaux de co(n)textualisation susmentionnés sont construits par l'analyse à partir d'une série de considérations théoriques ; de même, texte, corpus et outils sont des artefacts qui, de par ce statut commun, font de l'interprétation une construction de l'analyste. Mais l'interprétation, en étant construite par l'outil et par le lieu d'observation, n'est pas détachée du texte : elle est partie prenante du texte et du discours, intégrée au même processus de construction que ces derniers.

Bibliographie

Adam J.-M. (1999). Linguistique textuelle. Des genres de discours aux textes. Paris : Nathan.

Adam J.-M. (2006). « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse des données textuelles », Lexicometrica – actes JADT'2006,

Authier-Revuz J. (2001). « Le discours rapporté », in Thomassone R. (éd.) Une langue, le français. Paris : Hachette, 192-201.

Bakhtine M. (1984 [1934]). Esthétique de la création verbale. Paris : Gallimard.

Beaugrande R. de, Dressler W. (1981). Introduction to Text Linguistics. London / New York : Longman.

Charolles M. (1989). « Coherence as a Principle in the Regulation of Discursive Production », in Heydrich W., Neubauer, F., Petöfi, J. S., Sözer, E. (eds.), Connexity and Coherence. Analysis of Text and Discourse. Berlin : de Gruyter, 3-15.

Charolles M. (1995). « Cohérence, pertinence et intégration conceptuelle », in Lane Ph. (éd.), Des discours aux textes : modèles et analyses. Rouen : Publications des Universités de Rouen et du Havre, 39-74.
Cislaru G. (2008). « L'intersubjectivation comme source de sens : expression et description de la peur dans les écrits de signalement », Les Carnets du Cediscor 10 : 117-136.
Cislaru G., Pugnière F. et SitriF. (dir.) (2008). Les Carnets du Cediscor 10 (« Analyse de discours et demande sociale : le cas des écrits de signalement »). Paris : PSN.
Cislaru G., Sitri F. (2008). « La représentation du discours autre dans des signalements d'enfants en danger : une parole interprétée ? », Circulation des discours et liens sociaux. Le discours rapporté comme pratique sociale (5-7 octobre 2006, Université Laval). Québec : Editions Nota Bene.
Cornish F. (2006). « Relations de cohérence en discours : critères de reconnaissance, caractérisation et articulation cohésion-cohérence », Corela, Numéros spéciaux, Organisation des textes et cohérence des discours. Disponible en ligne à l'URL : http://edel.univ-poitiers.fr/corela/document.php?id=1280.
Garnier S. (2008). « L'évaluation dans les rapports de signalement », Les Carnets du Cediscor 10 : 79-91.
Guilhaumou J. (2002). « Le corpus en analyse de discours : perspective historique », Corpus 1 : 21-49.
Guilhaumou J., Maldidier D. (1979). « Courte critique pour une longue histoire. L'analyse du discours ou les (mal)heures de l'analogie », Dialectiques 26 : 7-23.
Legallois D. (2006). « L'hypertextualité et virtualité comme modes de la construction des discours et des connaissances », Pratiques 129-130 : 139-156.
Habert B., Nazarenko A., Salem A. (1997). Les linguistiques de corpus. Paris : Armand Colin.
Maingueneau D. (1991). L'Analyse du discours : introduction aux lectures de l'archive. Paris : Hachette Supérieur.
Maldidier D. (1990). L'inquiétude du discours, textes de M. Pêcheux. Paris : Edition des cendres.
Mayaffre D. (2002). « Les corpus réflexifs : entre architextualité et hypertextualité », Corpus 1:51-69.
Mayaffre D. (2007). « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? », in Rastier F. et Ballabriga M. (dir.), Corpus en Lettres et Sciences sociales. Toulouse : PU de Toulouse Le Mirail, 15-25.
Mazière F. (2005). L'Analyse du discours. Histoire et pratiques. Paris : PUF.
Moirand S. (2004). « L'impossible clôture des corpus médiatiques. La mise au jour des observables entre catégorisation et contextualisation », TRANEL 40 : 71-92.
Munchöw P. von (2001). Contribution à la construction d'une linguistique de discours comparative : entrées dans le journal télévisé français et allemand, Thèse pour le doctorat, Université Paris 3 Sorbonne nouvelle.
Rastier F. (2001). Arts et sciences du texte. Paris : PUF.
Rastier F. (2007). « Le corpus en questions », in Rastier F. et Ballabriga M. (dir.), Corpus en Lettres et Sciences sociales. Toulouse : PU de Toulouse Le Mirail, viii-xiii.
Rastier F., Pincemin B. (1999). « Des genres à l'intertexte », Cahiers de praxématique 33 : 83-111.
Rousseau P. (2007). Pratique des écrits et écriture des pratiques. Paris : L'Harmattan.
Sitri F. (2008). « Observer et évaluer dans les rapports éducatifs : de la représentation d'un dire singulier à la description d'une situation », Les Carnets du Cediscor 10 : 95-116.

Van de Velde R. G. (1989). « Man, Verbal Text, Inferencing, and Coherence », in Heydrich W., Neubauer F., Petöfi J.-S., Sözer E. (eds.) Connexity and Coherence. Analysis of Text and Discourse. Berlin : de Gruyter, 174-217.
Viehweger D. (1989). « Coherence – Interaction of Modules », in Heydrich W., Neubauer F., Petöfi J.-S., Sözer E. (eds.), Connexity and Coherence. Analysis of Text and Discourse. Berlin : de Gruyter, 256-274.