# ESSAI DE SYNTHÈSE

**Michel BALLABRIGA**
**CPST, Université de Toulouse 2**

Le consensus est très large, semble-t-il, sur *l'utilité*, à distinguer de l'utilisation, des *bases* numérisées qui permettent d'accéder à d'énormes archives consultables immédiatement, et notamment pour les recherches en Sciences Humaines et Sociales. Des voies s'ouvrent à la philologie numérique et les ressources de l'édition numérique, notamment par la constitution d'*appareils critiques* enrichissant considérablement et renouvelant l'apparat critique, contribueront à l'évolution et à la facilitation des études de toute sorte. C'est un *de facto* des nouvelles technologies qui livrent là une matière, une substance en extension rapide dont on ne peut faire quelque chose que par la création de *corpus* dont la constitution dépend de la tâche projetée et on ne peut faire quelque chose d'un corpus de quelque étendue que par l'entremise d'un *outil logiciel* de traitement, d'interrogation. La *tâche* se profile avec le corpus et les outils, nouvelles entités numériques qui doivent être rendues compatibles afin que le corpus offre *prise* à l'outil. Conformément à l'axiome saussurien - c'est le point de vue qui crée l'objet - le corpus semble pouvoir être défini *a minima* (noyau invariant) comme une collection de textes en fonction d'un objectif, d'une recherche.
On peut craindre une instrumentalisation du corpus (et du texte donc) au service du logiciel ; c'est la question des moyens et des fins, pas nouvelle certes, mais d'un autre poids puisque la *machine scientifique oraculaire* a parlé ; la question de l'ADN textuel et la volonté/possibilité de "faire parler les textes", fort stimulantes et suggestives, peuvent éveiller d'autres échos moins plaisants : "les textes vont se mettre à table" (!). Il faut aussi être prudent sur la question de l'attribution : comment distinguera-t-on l'excellent pastiche d'un écrivain qui a intégré, par "innutrition", la façon, le *ductus* d'un auteur ? C'est toute la question de l'imitation quasi-obligée dans la tradition littéraire. Comment distingue-t-on du non assumé sans analyse linguistique précise (cf. "Brutus est un homme honorable" dans le Jules César de Shakespeare), la polyphonie en général ?
Des questions, (davantage liées à l'ethos de la recherche), peuvent apparaître au moment des *choix* au fil des étapes : constitution du corpus (dépendant de l'objectif d'une tâche), codage, annotations, étiquetage, enrichissement etc. et constitution de logiciels d'interrogation des corpus qui reposent sur certaines postures (théoriques, méthodologiques, épistémologiques) qu'il convient de ne pas *naturaliser* même par la technique ("ça va de soi") car elles ont des incidences sur la recherche et ses résultats : ceux-ci sont bien sûr fonction d'une théorie linguistique (d'une grammaire si on veut, à *évaluer* ainsi que les *résultats*) que l'on peut "critiquer".
On peut avoir l'impression que l'outil est plus complexe que les résultats qui prennent valeur de confirmation du su, du pressenti ; on peut avoir le sentiment que le corpus (global) mange le texte (comme localité) ; même si on fait des retours nécessaires mais ponctuels au texte (ou mieux, contexte), *on peut pratiquer ce genre de recherches sans lire les textes du corpus*, contre-partie de la globalisation (inversement, la brièveté du texte rend ces analyses délicates : ce type de recherche a besoin d'un corpus étendu). Est-ce qu'un échantillonnage ne suffirait pas pour certains discours répétitifs, peu innovants, où le figement et le stéréotype sont importants et presque de mise, pour formuler des interprétations, quitte à confirmer et à affiner par le reste du corpus ? Une bonne connaissance du genre permet peut-être une certaine économie descriptive logicielle. Tous les textes ne sont probablement pas justiciables d'une même méthode.
L'interprétation s'est parfois déplacée : corpus codé----> analyses automatiques----> résultats----> commentaires ; *on interprète des résultats (indirects) produits par l'assistance logicielle*, un autre texte, second ; l'interprétation apparaît aussi entre les résultats et les commentaires. *Cette phase interprétative ne paraît guère spécifiée* ; d'où, peut-être, le désir - scrupule tout à fait honorable - de retarder le plus possible et de contrôler l'interprétation - qui apparaît comme un principe de plaisir - en donnant la part la plus importante à la description - qui participerait du principe de réalité - et la mention (avec quelque regret) du nécessaire saut interprétatif, le plus tard possible... En fait, peut-être que la phase descriptive est plus "confortable", on ne dit pas plus facile, et les principes sont peut-être inversés... Au cœur des questionnements se trouvent notamment la question de *l'unité* et celle du *changement d'échelle* dont découle la question des rapports entre *micro-analyse et macro-analyse.*

Du type *d'unité* et de sa *catégorisation* - dans les recherches actuelles unités lexicales et parties du discours sont privilégiées dans la conduite de l'interrogation - dépend en grande partie le cadre méthodologique, mais le type d'unité et la catégorisation dépendent de la théorie et de l'épistémologie explicites ou implicites.
Le recours au mot (graphique) comme approximation incontournable est recevable, notamment en l'état actuel des techniques logicielles. La question du figement (de ses degrés) - le mot est une unité figée - et des segments répétés permet de complexifier cette question de *l'unité* (phraséologies, prêt-à-dire relevant de formes de la doxa) et de la *sémiose.* Ces nouvelles méthodes font aussi apercevoir de nouvelles formes insoupçonnées du fait du changement d'échelle et de la rapidité du parcours des associations attestées (en gros le télescope "remplacerait" le microscope) que l'humain seul ne peut réaliser. Ces outils ouvrent incontestablement la voie, entre autres, à une nécessaire *sémantique textuelle historique et comparée*, composante essentielle d'une *sémiotique des cultures*, notamment en permettant des études thématiques et topiques d'envergure, jamais permises jusque là. Il se produit certes une modification importante dans la perception des textes ; de nouveaux objets, de nouvelles unités apparaissent, ainsi que de nouvelles façons d'interpréter. Comme le signalait le texte d'orientation, un nouveau rapport à l'empirique se constitue, susceptible d'entraîner de nouvelles formes d'élaboration des connaissances, mais il convient au plus haut point de ne pas enfermer la démarche dans un seul point de vue méthodologique et théorique, de ne pas le *naturaliser*, et de développer par des voies et instruments spécifiques les aspects plus proprement sémantiques : "L'enjeu consiste à passer du *zero meaning* (chaîne de signifiants avec traitement statistique) à l'analyse thématique, à pallier l'absence de "données sémantiques" en tirant profit de la théorie sémantique" (F. Rastier, 2001, *Arts et Sciences du Texte*, P.U.F., p. 206).
Pour le dire grossièrement, le contenu des mots est constitué de mots et cette configuration interne, sorte de sédimentation au niveau de la langue, est *a priori* variable en contexte du fait d'une dynamique propre de l'échange textuel par actualisation ou inhibition notamment : "L'analyse en traits sémantiques reste cruciale, car deux occurrences d'un topos qui n'ont aucune lexicalisation en commun doivent pouvoir être reconnues ; c'est d'ailleurs le seul moyen de sortir de la logique documentaire du mot-clé" (*ibid*. p. 218)". Ceci nous renvoie aux modes d'existence divers d'une unité : on connaît l'exemple des étudiants pensant avoir entendu le mot *avalanche* dans la phrase : "la neige dévalait furieusement la pente" ; cela est à traiter non pas, de façon réductrice, comme une illusion mais par une approche complexe de la perception sensorielle, mentale et sémantique (cf. aussi *ibid*. p. 200-201 : "alors que ce mot [ennui] se rencontre seulement quatre fois dans Madame Bovary, les composants du thème apparaissent souvent [...] Le mot *ennui* est absent, mais les sèmes caractéristiques du thème de l'Ennui se répètent massivement"). Cela conduit bien sûr aux problèmes philologiques et herméneutiques liés à la numérisation et au traitement informatique et ces phénomènes ne sauraient être évacués (puis niés) parce que la technique actuelle ne sait pas encore les traiter convenablement... L'association de termes sur une *surface* impressionnante et complexe (les *cooccurrents*) ne doit pas masquer la question (perceptive et sémantique) de ce qui se passe, du point de vue de la *profondeur* et du *volume*, sous les mots, dans les mots, entre les mots (les *corrélats*). D'où les travaux (en cours) pour des annotations d'un autre type (plus sémantique) : les isotopies et les thèmes, de nature *sémique*, se distribuent sur des termes qui peuvent être *catégoriellement différents*, deux caractéristiques qui posent problème en lexicométrie. Il est possible que la prise en compte des unités lexicales soit (nécessaire et) suffisante pour certains corpus, qui d'ailleurs peuvent s'accommoder d'un simple balayage. Se pose la question de l'adéquation de la méthode, et de ses phases, à différents discours, ce qui renvoie à leur mode de production. L'analyse en traits sémantiques est-elle à réserver au discours littéraire notamment ? Rien n'est moins sûr : elle est utile pour des discours où le stade lexical *paraît* suffisant, c'est-à-dire en fait qu'on ne peut aller au-delà de ce que permet la méthodologie employée, ce que rendrait possible une analyse qualitative : "le quantitatif et le qualitatif ne s'opposent aucunement: seule une analyse qualitative peut rendre significatifs des phénomènes quantitatifs remarquables" (*ibid*. p. 214). La prise en compte du palier lexical est dans tous les cas nécessaire et il faut probablement prévoir et des *stades* interprétatifs (lexical, sémique, que l'on n'opposera pas, etc.) qui ne sont peut-être pas tous pertinents pour tous les genres de textes, de corpus, et des *phases* dans la tâche interprétative assistée, où se repose la question des moyens et des fins. De même, l'intérêt des macro-analyses ne périme pas celui des micro-analyses (par exemple sur *un* texte, localité qui peut être considérée comme une

globalité) qui demeurent nécessaires dans certaines tâches ou parties de tâches, les résultats des études globales permettant d'ailleurs de mieux fonder et d'enrichir les analyses locales, sans oublier la question du plaisir du texte : le "microscope" demeure donc fort utile.
Le logiciel, qui peut être une fin dans une pratique (création de logiciels), apparaît pour la majorité des pratiques utilisatrices comme un *pouvoir-faire* au sens sémiotique, d'une grande potentialité descriptive. Reste à spécifier ce faire : interprétatif, probatoire, heuristique ; dans ce dernier cas, il peut servir à interroger les textes ; mais de quoi dépend la forme de ces interrogations, qu'est-ce qui les suscite, les oriente ? L'interrogation est au centre d'un faisceau dont les extrémités sont reliées à l'objet, à l'objectif, à la théorie, à l'outil (sans préjuger des relations secondes outil-théorie etc.). Ces faires sont d'ailleurs probablement en interaction. Des choix théoriques président à l'étiquetage, au traitement, à l'analyse, à l'interprétation ; il convient de ne pas en faire des absolus descriptifs, de les naturaliser (même si on reconnaît des variations dans les classifications, mais en restant toutefois dans une théorie), de voir que d'autres descriptions sont possibles à partir d'autres postures théoriques. Ce relativisme, *situé*, s'oppose au dogmatisme et il gagne aussi la notion de corpus qui "présuppose une préconception des applications envisagées" (F. Rastier lors d'une intervention à ce colloque) et induite notamment par des positionnements théoriques et/ou méthodologiques : "L'informatique n'est pas un organon théorique, et son usage ne préjuge en rien le bien-fondé d'une thématique assistée" (*ibid*. p. 214).
Il a été largement question des documents et outils numériques, mais la problématique de l'interprétation demeure centrale et ne saurait se détacher d'abord de certaines questions : qu'est-ce qu'interpréter, qui interprète, quoi interpréter, comment interpréter, pour qui et pour quoi ?
Il convient de ne pas opposer brutalement description et interprétation. Où commence l'interprétation, où s'arrête la description ? Quelle est la part interprétative de la description ? Le *versus* entre ces deux termes n'est peut-être pas de mise dans ce continu à seuil dynamique ; on ne décrit pas sans appuis interprétatifs. Il a été dit dans ce colloque, dans une certaine intention : "la catégorisation, c'est l'interprétation" ; on pourrait aussi bien dire : "la catégorisation, c'est *de* l'interprétation"
Il est patent que l'interprétation - *le fait d'interpréter* - est bien l'affaire de tous. Toutefois, l'interprétation - comme *processus scientifique raisonné* - met en oeuvre une méthodologie appuyée sur une théorie aux bases épistémologiques clarifiées ; étudiant de manière critique les conditions de possibilité de l'interprétation, ses résultats ne sont pas simplement "affaire d'interprétation" au sens obvie. Cet aspect-là, compris ainsi, a été peu présent, sinon représenté.
La considération des *moyens* et des *fins* et de leur *dialectique* semble vraiment cruciale dans la thématique de ce colloque :
- la question des *pratiques sociales* en général paraît primordiale (c'est le "pour qui ?") : si le corpus ainsi que son exploitation dépendent de l'objectif, celui-ci est "validé" par la pratique sociale (ou le niveau de pratique sociale) d'où l'activité part, du fait de quelque initiative, et où elle retourne : les types de corpus sont relatifs aux pratiques sociales et aux tâches ; dans une visée esthétique une œuvre intégrale peut constituer un corpus. Il convient d'avoir des relations indispensables avec les acteurs de la pratique pour ne pas risquer l'involution : la pratique sociale commande. Cette relativisation peut guider la pertinence des tâches entreprises (construction de dictionnaires, grammaires, études de langues de spécialités, études sémantiques...)
- on peut souhaiter une diversification selon les intérêts des utilisateurs et la nature du corpus ; l'instrument informatique, quand il est conçu et utilisé comme outil ou boîte à outils (dans une conception noble du bricolage) serait à modifier en conséquence selon les tâches et parties de tâches - dans des formes d'appropriation, voire de détournement de l'outil - en relation avec une adaptation théorique nécessaire : en changeant de fond applicatif, la théorie comme *forme* peut se *transposer* (cf. la "dégradation" théorique, assez vilainement dénommée).
Enfin, sur cette question du corpus et des documents numériques, le colloque a été un lieu de rencontre, pouvant dynamiser un vrai projet fédérateur indispensable en Lettres et Sciences Sociales. Le *relativisme situé* qui a été évoqué devrait avoir comme pendant des possibilités réelles et urgentes *d'interdisciplinarité* - statut du corpus dans les différentes disciplines ; méthodes et réflexions différentes ; comparaison des points de vue ; partage des sources, croisement des données et mutualisation pour les chercheurs en Sciences Sociales et Humaines - qui/qu'ouvrirait, constituée d'objectifs communs, une zone *transdisciplinaire* consistante, afin de se donner les moyens de penser, *culturellement*, la complexité culturelle.