Questions vives

Questions Vives
Recherches en éducation

N° 28 | 2017
De l'indifférenciation à la différenciation

Corpus numériques et accès à la culture

François Rastier

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/questionsvives/2421
DOI : 10.4000/questionsvives.2421
ISSN : 1775-433X

Éditeur
Université Aix-Marseille (AMU)

Édition imprimée
Date de publication : 29 décembre 2017
ISBN : 978-2-912643-52-0
ISSN : 1635-4079

Référence électronique
François Rastier, « Corpus numériques et accès à la culture », *Questions Vives* [En ligne], N° 28 | 2017, mis en ligne le 06 novembre 2018, consulté le 02 mai 2019. URL : http://journals.openedition.org/questionsvives/2421 ; DOI : 10.4000/questionsvives.2421

Corpus numériques et accès à la culture

François Rastier

1. Défis épistémologiques

1 *L'accès aux documents numériques* – La didactique et les sciences de l'éducation sont affrontées aux outils informatiques – mais ces outils se sont tellement banalisés que le numérique a rejoint l'électrique dans le train-train quotidien. Si leur usage fait encore débat, il ne nous retiendra pas ici. Nous questionnerons l'accès aux objets culturels qui passe à présent pour l'essentiel par les moteurs de recherche et les services grand public comme Google Images ou YouTube. Deux questions nous retiennent.

2 1/ Il nous semble nécessaire de rappeler et de souligner la distinction jadis évidente mais à présent obscurcie entre l'objet et le document qui le représente. Par exemple, à la requête « Van Gogh », Google Images renvoie des images de formats comparables, aux couleurs sursaturées : les seules informations immédiatement disponibles sont le lieu d'extraction (éditeur), le nombre de pixels. Disparaissent les titres, les dates, les lieux d'exposition, les formats, les reliefs pourtant caractéristiques de la matière picturale, et chaque tableau devient une sorte de vignette commémorative.

3 2/ Alors que les corpus d'interprétation ont un rôle essentiel, chaque tableau se trouve plongé dans un ensemble hétérogène qui ne permet pas de le contextualiser. Apparaissent en effet, sur le même plan et sans rien qui permette de les distinguer, des photos d'œuvres authentiques, des photos prises avec des logiciels *arty* qui « font » du Van Gogh, des couvertures de livres, des images de baskets en toile, de papiers peints, de la stèle funéraire du peintre, etc.

4 De fait, les indexations des moteurs de recherche généralistes ne font que refléter sans contrôle les associations aléatoires effectuées par des utilisateurs. Soit, mais les sites qui se prétendent à une vocation culturelle ne font pas autrement. Par exemple, pour le dessin, Google Arts & Culture proposera un parcours de recherche en fonction de la couleur et de la chronologie, comme le vert ou le XVII^{e} siècle. Ses catégories descriptives

mêlent les types de support (Vergé, Vélin), les modes d'inscription (Graphite, Aquarelle, Encre) et les instruments (comme Stylo). Sous la rubrique Vergé, on trouve une gravure au burin de Dürer, qui n'est aucunement un dessin. Sous la rubrique Stylo, on trouve un dessin de Dürer (*Pen and brown ink, brown wash, corrections in white gouache*), inutile de préciser que le stylo n'existait pas à l'époque de Dürer. Les œuvres de techniques mixtes sont rangées aléatoirement sous une seule catégorie, fût-elle inepte. Sous la même rubrique Stylo, on trouve aussi une encre de Van Dijk, une aquarelle de William Blake, etc. Ainsi substituées à tout cadre de compréhension et d'interprétation, les métadonnées deviennent un obstacle à la connaissance, remplacée par l'accès à des documents-substituts.

5 Les données n'ont pas de sens. Pour qu'elles puissent en acquérir, il faut d'une part avoir une vue claire de leur provenance et de leur élaboration, selon le principe que dans les sciences de la culture comme dans les autres, les données, c'est ce qu'on se donne. En d'autres termes, les données sont la première concrétisation d'une hypothèse, celle qui permet de les circonscrire. Il faudra ensuite, pour leur conférer un sens, détailler pas à pas, de manière réflexive, une méthodologie qui permettra d'infirmer ou de confirmer l'hypothèse première. On ne trouve pas toujours ce que l'on cherche, on trouve souvent ce qu'on ne cherchait pas, mais on ne peut rien trouver par la simple manipulation de données non qualifiées.

6 On a pu dire que l'homme cultivé était aujourd'hui celui qui ne cliquait pas sur le premier lien affiché en haut de première page de résultats. L'amateur fervent ira peut-être jusqu'à la page deux. Seul l'érudit prodigue de son temps consultera les documents originaux, pour autant qu'ils restent accessibles.

7 On donnait jadis à l'école la mission d'accéder à la culture et de compenser les inégalités sociales sur ce point. Cette mission demeure certes, mais l'accès à la culture pourrait bien consister désormais à savoir se *protéger* des flux de « données » hétérogènes aimablement déversés par des algorithmes opaques. Cette déontologie a évidemment des enjeux éducatifs – et démocratiques – pour éclairer la maîtrise personnelle des informations et des connaissances.

En précisant les principes déontologiques d'une sémiotique de corpus, nous allons donc formuler les conditions d'une contextualisation réfléchie, nécessaire à l'interprétation des objets culturels.

8 *Corpus et méthodologie des sciences de la culture* – La méthodologie historique et comparative réunit les sciences de la culture depuis leur formation au début du XIX^{e} siècle, notamment en linguistique (de Bopp à Saussure, Benveniste, Coseriu, De Mauro, etc.) ; en anthropologie (de Boas à Lévi-Strauss) ; en histoire et science des mythologies et des religions (Dumézil), en iconologie (de Warburg à Panofsky), etc.

9 La sémiotique de corpus participe du programme comparatiste. Également issue du projet comparatiste par son épistémologie comme par sa méthodologie, la sémiotique d'inspiration saussurienne réunit les conditions pour participer à ce projet de caractérisation contrastive qui a l'ambition de restituer la complexité et la diversité interne des objets culturels. Cependant, (i) la dimension appliquée et l'instrumentation raisonnée exigent des clarifications méthodologiques ; (ii) le choix du corpus impose une perspective critique ; (iii) le choix des tests instrumentés redouble les problèmes herméneutiques ; (iv) l'annotation (métadonnées et balisage) renoue avec la problématique philologique.

10 Depuis deux siècles, les corpus des sciences de la culture ont été élaborés par un patient travail de déchiffrement des écritures disparues, de documentation des monuments, de collections scientifiques, des relevés archéologiques, de l'épigraphie, de la diplomatique, des carnets de recherche des ethnologues, des recueils des linguistes, des historiens de l'art, des religions, de la mythographie, etc. De nombreuses collectivités sont de longue date engagées dans une réflexion sur la numérisation et l'analyse assistée des documents : outre bien entendu les sciences de l'information, il faut mentionner entre autres l'histoire, la sociologie, la linguistique, l'archéologie, les études littéraires.

Toutes les disciplines ont maintenant affaire à des documents numériques et cela engage pour elles un nouveau rapport à l'empirique. Les nouveaux modes d'accès aux documents engagent-ils de nouvelles formes d'élaboration des connaissances ?

2. Pour approfondir le concept de corpus

11 *Définition* – Un auteur célèbre définissait le corpus comme un « vaste ensemble de mots ». Cependant l'objet empirique de la linguistique est fait de textes (oraux ou écrits), non de mots ou de phrases – qui ne s'observent pas à l'état isolé, ou du moins appartiennent toujours à un genre et un discours. Si le signe est l'unité élémentaire, la performance sémiotique est pour une sémiotique évoluée l'unité *minimale* et le corpus l'ensemble dans lequel cette unité prend son sens.

12 Commençons par une définition positive. Un corpus est un regroupement structuré de performances sémiotiques intégrales, documentées, éventuellement enrichies par des étiquetages, et rassemblées : (i) de manière théorique réflexive en tenant compte des pratiques sociales et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.

13 Tout corpus suppose en effet une préconception des applications pour lesquelles il est rassemblé : il dépend étroitement du point de vue qui a présidé à sa constitution.

De fait, tout regroupement de documents ne mérite pas le nom de corpus. Ainsi, une collection documentaire peut regrouper des documents numériques de statuts divers, sans qu'aucun critère sémiotique ne permette leur totalisation. Même organisée en base de données, une collection documentaire ne devient pas pour autant un corpus.

14 Dans la tradition, la notion de corpus a d'abord été définie de manière canonique dans les domaines religieux, juridique, voire littéraire. À cette conception canonique, on semble préférer aujourd'hui une conception éclectique. Cependant, un corpus n'est pas plus un sac de mots qu'un nébuleux intertexte.

15 La conception *documentaire* n'en retient que des variables globales caractérisant les documents, sans tenir compte de la complexité de leur structure. Dans cette conception logico-grammaticale, le corpus se résume à un échantillon de la langue, un réservoir d'exemples ou d'attestations. En revanche, la conception *philologique-herméneutique* tient compte des rapports entre performances, selon le principe critique qui est traditionnellement celui de la philologie, devenue ici numérique.

16 Pour étudier une performance, le « bon corpus » est d'abord constitué des performances qui partagent le même genre. Un *champ générique* est un groupe de genres qui contrastent voire rivalisent dans une pratique sociale : par exemple, au sein du domaine photographique de la presse *people*, le *portrait* se différencie de l'*indiscrétion* (non posée, pixellisée, prise au téléobjectif).

17 Il semble utile de distinguer quatre niveaux. (i) L'*archive* réunit l'ensemble des documents accessibles pour une tâche de description ou une application. Elle n'est pas un corpus, parce qu'elle n'est pas constituée pour une recherche déterminée. (ii) Le *corpus de référence* est constitué par un ensemble de documents, sur lequel on va contraster les corpus d'étude. (iii) Le *corpus d'étude* est délimité par les besoins de l'application. (iv) Enfin, les *sous-corpus de travail* varient selon les phases de l'étude et peuvent ne contenir que des passages pertinents des performances sémiotiques étudiées[1].

18 *Une rupture nécessaire* – Les fonctionnalités statistiques de base ont été mises au point dès les années 1960 (application du test de l'écart réduit, analyses factorielles, classification automatique), et l'essentiel était acquis dans le courant des années 1980 (comme les calculs de spécificités ou de cooccurrences).

19 Beaucoup cependant reste à faire pour convaincre de la nécessité de travailler sur corpus. La technicité, le détour instrumental, la notion même de méthode expérimentale, inquiètent certains ; l'attachement à la recherche sans sanctions empiriques, parfois même dans des disciplines littéraires la répugnance à l'égard de toute objectivation censée porter atteinte à la subjectivité souveraine des auteurs et des lecteurs, tout cela conduit certains à considérer l'étude des corpus comme un leurre. Ils formulent une objection récurrente : on ne trouve jamais que ce que l'on cherche. Soit ils regrettent par là que l'on vérifie l'intuition sans songer qu'il est parfois difficile de prouver des évidences, ni que cela fait partie de l'ingrate mission des sciences ; soit encore ils estiment qu'on trouve toujours quelque chose : c'est faux, car des résultats bruités peuvent inviter au silence.

20 De fait, on ne trouve pas toujours ce que l'on cherche, mais souvent autre chose que l'on ne cherchait pas : de nouveaux observables. Certes, on ne trouve trop souvent que ce que l'on sait voir et l'on reste dépendant d'un état de l'art et des problématiques routinières de la « science normale » ; une démarche critique permet cependant de les dépasser ensemble.

21 *Le caractère critique de l'expérimentation.* – Le doute positif relève de l'attitude critique nécessaire à toute problématisation scientifique, mais les besoins méthodologiques restent d'autant plus grands que les sémiotiques dont on dispose sont pour l'essentiel des sémiotiques du signe, à postulats mentalistes (sémiotique cognitive, théorie du prototype, etc.) et sans protocoles expérimentaux. La sémiotique de corpus peut cependant faire avancer la réflexion sémantique au niveau méthodologique (pratique) comme au niveau épistémologique (théorique).

a) Le sens étant fait de différences, le détour méthodologique par l'instrumentation permet de construire des différences : entre passages, entre performances, entre auteurs, genres et « discours ». La pertinence n'émerge pas du quantitatif, mais de la rencontre entre deux horizons : la pertinence « subjective » déterminée par la tâche et la pertinence « objective » propre aux inégalités qualitatives au sein des performances sémiotiques et entre elles.

b) Au niveau épistémologique, le détour expérimental permet l'objectivation : (i) en infirmant ou en confirmant des hypothèses, (ii) en faisant ressortir les régularités structurelles de l'objet, quand diverses procédures instrumentales parviennent à des résultats concordants malgré les différences de matériau expérimental, d'échelle, etc.

22 À la classique dualité entre induction et déduction dans les disciplines d'observation, le renouvellement méthodologique favorisé par les corpus numériques engage à substituer

le cycle suivant : (i) analyse de la tâche et production des hypothèses ; (ii) constitution d'une archive et sélection d'un corpus de référence ; (ii) élaboration des corpus de travail ; (iii) traitement instrumenté de ces corpus, en contrastant corpus de travail et corpus de référence ; (iv) interprétation des résultats et retour aux sources textuelles pour valider l'interprétation. La puissance propre de ce dispositif heuristique permet de faire émerger de *nouveaux observables* inaccessibles autrement : par exemple, la phonostylistique, jadis condamnée à l'intuition, se voit à présent pourvue de moyens d'investigation par des statistiques sur corpus phonétisés. En outre, l'utilisation d'une instrumentation scientifique (analyseurs, étiqueteurs, etc.) participe du processus d'objectivation : les objets culturels ont beau dépendre de leurs conditions d'élaboration et d'interprétation, les valeurs qu'ils concrétisent peuvent cependant être objectivées comme des faits.

23 La sémiotique de corpus pourvoit ainsi la sémiotique d'un domaine où elle peut élaborer des instruments et définir une méthode expérimentale propre : elle ouvre aussi des champs d'application nouveaux et engage un mode spécifique d'articulation entre théorie et pratique. Sans renoncer à l'élaboration théorique, elle en limite la portée aux corpus étudiés, et, sans se satisfaire de la seule démarche déductive, procède par essais et erreurs. En bref, la recherche part d'une diversité constatée, l'unifie dans le point de vue qui préside à la collection du corpus, éprouve enfin son objectivité par l'investigation instrumentée.

3. Redéfinir le concept de système de signes

24 On a trop souvent réduit les sémiotiques à des inventaires de signes et de règles. Il faut cependant tenir compte, outre du *système*, des *corpus* (corpus de travail et corpus de référence), de l'*archive* (historique), enfin des *pratiques* sociales où s'effectuent les activités sémiotiques. Pour l'essentiel, une sémiotique repose sur la dualité entre un *système* (condition nécessaire mais non suffisante pour produire et interpréter des performances) et des *corpus* de performances documentées.

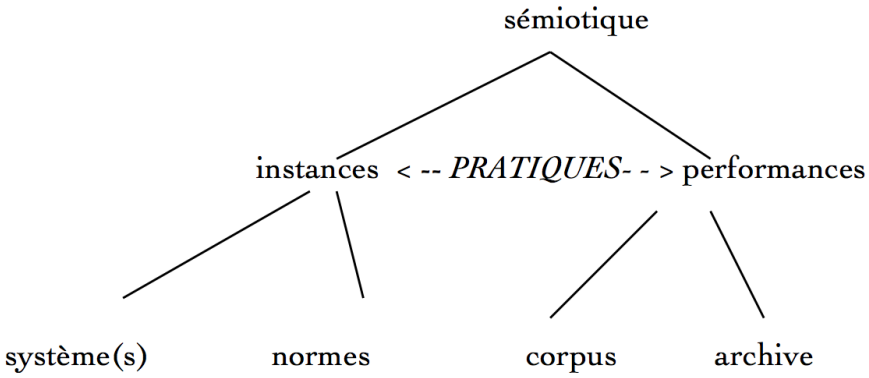
25 Non contradictoire, la dualité dynamique entre corpus et système constitue la sémiotique dans son histoire. En évoquant les corpus et non les signes, nous soulignons qu'une sémiotique n'est pas un système de signes – comme le serait un code. Saussure, à qui l'on prête cette définition, ne l'a d'ailleurs jamais formulée. Un signe au demeurant n'a pas de définition intrinsèque : il n'est qu'un *passage* ou zone de localité, certes réduit, d'une ou plusieurs performances sémiotiques auxquelles il renvoie.

26 Entre le corpus et le système, les normes assurent un rôle de médiation : ancrées dans les pratiques sociales, les normes de genre et de style témoignent de l'incidence des pratiques sociales sur les performances qui en relèvent[2]. Il paraît donc préférable de considérer qu'un système sémiotique complexe ne se limite pas à des règles, mais comprend aussi des *normes* diversement impératives. Les règles et les normes ne diffèrent sans doute que par leur régime d'évolution diachronique. Par ailleurs, en synchronie, toute règle voisine avec des normes qui accompagnent voire conditionnent son application. Ainsi, à la différence d'un langage formel, un système sémiotique complexe se décline en régimes structurels différents selon les niveaux et paliers d'analyse. Ses domaines d'organisation locaux ou régionaux ne sont pas unifiés dans une hiérarchie attestant l'existence d'un système unique et homogène, comme en témoigne au

demeurant l'évolution continue des langues qui trouvent dans leur hétérogénéité systémique le moteur interne de leur changement perpétuel par perturbations et ajustements.

27 Non moins plurielles que les instances, les performances se spécifient *a minima* par la distinction entre les corpus (de travail et de référence) et l'archive. À la grande diversité des pratiques sociales correspond celle des corpus produits en leur sein. Soit, schématiquement :

Figure 1 : Instances et performances



Le schéma ci-dessus jouit d'une grande généralité et peut convenir à des sémiotiques non verbales complexes, comme l'iconologie, par exemple. Nous l'avons d'ailleurs utilisé pour un système d'aide à l'indexation de photographies.

28 La dualité entre *langue* et *parole* chez Saussure est un cas particulier du rapport entre instances et performances. Au plan méthodologique, la flèche qui va des performances aux instances symbolise l'extraction de régularités ; et la flèche inverse symbolise la caractérisation de singularités, les deux processus restant interdépendants.

29 Enfin, au plan épistémologique, il est vraisemblable que la dualité entre instances et performances (ou système(s) et corpus) traduit une dualité de problématiques, l'une de tradition logique et grammaticale, l'autre de tradition rhétorique et/ou herméneutique. La problématique logico-grammaticale privilégie les instances (car elle s'appuie sur une ontologie), alors que la problématique rhétorico-herméneutique privilégie les performances, car elle repose sur une praxéologie. Dans l'histoire des réflexions sémiotiques occidentales, tributaires de la problématique logico-grammaticale, les instances dominent les performances : de la théorie scolastique du langage comme faculté qui s'effectue par des actes (contenus en puissance dans la faculté), on en est par exemple venu à la théorie chomskyenne de la générativité à partir de règles.

4. Pour une conception non antinomique de la dualité entre instances et performances

30 *Paliers, niveaux et remembrement* – À cause de la séparation infondée entre syntaxe, sémantique et pragmatique[3], une double séparation s'est établie, de fait et non de droit, entre les domaines de la sémiotique. Elle peine à concevoir le rapport du local au global, médiatisé par plusieurs paliers d'organisation de complexité croissante. Même, la relation entre signifié et signifiant qui constitue le signe ne va aucunement de soi, n'a rien

d'inconditionné et reste médiatisée par les structures des performances, considérées tant au plan du contenu qu'au plan de l'expression. Or, précisément, la sémiotique de corpus permet l'étude du rapport entre global et local tant au plan du contenu que de l'expression, considérés séparément et dans leur interrelation constitutive de la sémiosis textuelle. Les complémentarités entre paliers de complexité sont illustrées par des phénomènes de *solidarité d'échelle* qui en sémiotique ne sont pas étudiés jusqu'à présent, faute de moyens techniques pour les cerner. Un peu comme l'imagerie cérébrale pour les neurosciences, les instruments de la sémiotique de corpus permettent à présent d'étendre le champ des observables.

31 *Les corrélations entre plans du contenu et de l'expression* – Elles sont cruciales pour la sémiotique des performances, car elles permettent d'aborder la question de la sémiosis. Par exemple, au plan *graphique*, alors que la ponctuation n'est pas considérée comme sémantique et qu'elle est tout simplement absente des grammaires formelles, l'étude en corpus permet de souligner les corrélations entre contenus lexicaux et ponctèmes. Ainsi, dans un corpus romanesque, Évelyne Bourion (2001) a ainsi pu confirmer la corrélation entre des noms de sentiments et les ponctuations dans les contextes où ces noms apparaissent. Ainsi les sentiments ponctuels, brusques, comme la *colère* ou la *joie* sont-ils fortement associés aux points de suspension.

32 Les corrélations entre plans du contenu et de l'expression rendent licite la notion de *contextualité hétéroplane* : le contexte d'une unité sur un plan, expression ou contenu, est constitué par d'autres unités sur le même plan, mais aussi sur l'autre. On ressent le besoin d'une théorie qui puisse penser ces corrélations, c'est-à-dire d'une linguistique informée par une sémiotique textuelle.

33 Les mêmes types de corrélations sont toutefois à l'œuvre dans des corpus non linguistiques. Ainsi le projet européen *Princip.net* de détection automatique de sites racistes (cf. l'auteur, 2011, ch. 7) a mis à profit des critères de « bas niveau » comme la ponctuation (un antiraciste ne redouble jamais un point d'exclamation), la casse (un antiraciste n'écrit jamais une phrase en majuscules), les polices de caractères, voire les codes html (les images sont caractéristiques des sites racistes).

34 Les corrélations entre plans du contenu et de l'expression ont aussi un enjeu immédiat pour les applications comme la catégorisation de documents, la détection automatique de sites, etc. En pratique, elles permettent, dès lors que la catégorisation des documents du corpus d'apprentissage tient compte d'une classification évoluée, d'éviter des traitements sémantiques complexes et aléatoires.

35 Les méthodes de recueil et d'analyse de corpus ainsi mises au point s'adaptent aussi à d'autres sémiotiques, où le repérage de corrélations entre contenu et expression permet par exemple l'identification rapide des genres. Ainsi dans une application d'assistance à l'indexation d'images fixes (projet Semindex), les genres d'un corpus de presse *people* ont-ils pu être discriminés en fonction de critères d'expression : par exemple, toute photo à gros grain peut être classée dans le genre de l'*indiscrétion* – qui suppose l'usage d'un téléobjectif.

36 *Vers un remembrement des disciplines* – En traitant les corpus, la sémiotique renoue nécessairement avec les performances complexes, donc avec la philologie et avec l'herméneutique : la philologie pour les établir et les documenter, l'herméneutique pour les interpréter.

37 L'essor de la sémiotique de corpus conduit notamment à préciser le rapport entre performances sémiotiques et documents, d'autant plus qu'en perdant son unicité, le document numérique se dépouille des qualités du document unique de l'archiviste : authentifiable, doué par sa continuité matérielle d'une intégrité (même quand il est fragmentaire), non reproductible, il pouvait faire autorité. À présent, l'affichage par pixels détruit toute continuité matérielle qui empêchait les falsifications. Alors qu'une critique initiale suffisait à établir le document, il faut à présent une critique continue pour maintenir sa fiabilité. L'établissement des significations doit souvent passer par une succession de versions, dont chacune est le support et le résultat d'une opération de lecture. En changeant ainsi de régime, l'objectivation peut progresser sans pouvoir jamais être considérée comme établie, ce qui engage à rompre avec l'objectivisme pour promouvoir une objectivation critique indéfinie.

38 Toutefois, ce que le document perd en stabilité, il le gagne en biais d'interrogation. Les logiciels appellent une réflexion théorique sur l'étiquetage, sur les rapports entre méthodes qualitatives et quantitatives : on peut par exemple croiser les résultats de plusieurs méthodes pour faire apparaître de *nouveaux observables*. C'est autant aux sémioticiens qu'aux informaticiens de faire des propositions sur ce point : pour aborder ces questions, la voie méthodologique et la voie épistémologique n'ont rien de contradictoire – d'autant plus que l'informatique est elle-même une technologie sémiotique...

5. Sémiotique des performances

39 *Quelques principes* – Une performance sémiotique isolée n'a pas plus d'existence que le mot ou la phrase isolés : pour être produite et comprise, elle doit être rapportée à un genre et à une pratique sociale.

40 Les corpus ne sont pas simplement des réservoirs d'attestations, ni même des recueils de performances inscrites sur des supports. Dès lors qu'ils sont constitués de façon critique, en s'entourant des indispensables garanties philologiques, ils peuvent devenir le lieu de description des trois régimes de sémioticité : génétique, mimétique, herméneutique. Une performance en effet trouve ses sources dans un corpus, elle est produite à partir de ce corpus et doit y être maintenue ou replongée pour être correctement interprétée : le régime génétique et le régime herméneutique se règlent ainsi l'un sur l'autre. Quant au régime mimétique, il dépend aussi du corpus et notamment de la doxa dont il témoigne.

41 Si l'on convient de ces constats, il faut encore, pour les rendre opératoires, déterminer les grandeurs ou « unités » sémiotiques et caractériser leurs relations au sein de la performance et entre performances, en fonction des parcours qui structurent dynamiquement le corpus et justifient *a posteriori* sa constitution. Le programme intellectuel de la sémiotique interprétative conduit ainsi à un remembrement de la sémiotique autour du concept de performance, ce qui engage à renouer avec des formes nouvelles de la philologie et de l'herméneutique[4].

42 *Les documents et la reconquête de l'expression* – Dès lors que le format des documents qui composent le corpus est défini de manière critique, des méthodes comparables peuvent s'appliquer à des textes, à des images, à des musiques, etc. Cela ouvre de grandes perspectives, non seulement pour l'unification de la sémiotique, trop souvent divisée en

domaines sensoriels (vision, audition, etc.), mais encore pour l'analyse des sémiotiques complexes et syncrétiques qui sont le principal objet empirique de la discipline.

43 Plutôt que de discuter abstraitement des disciplines, il nous paraît plus utile de détailler les relations entre le *document*, qui relève pour l'essentiel de la philologie, la performance, qui relève (ou devrait relever) de la sémiotique, et l'*œuvre*, qui relève plus particulièrement de l'herméneutique dans la mesure où elle appelle une interprétation critique pour l'aborder dans sa complexité. Il faut pour cela formuler un modèle sémiotique des performances et des objets culturels qui articule non seulement le contenu et l'expression, mais aussi les pôles du Point de vue (concept herméneutique) et de la Garantie (concept philologique – cf. l'auteur, 2018). Cela conduit à poser des questions de valeur et de légitimité : c'est par la médiation d'une sémiotique étendue à ces questions que l'herméneutique (trop idéalisée) et la philologie (trop positivisée) pourraient se rencontrer dans une situation nouvelle, ouverte par l'essor de la sémiotique de corpus. Détaillons ce point.

44 (i) On préfère souvent à présent traiter des performances en termes de documents, en privilégiant la conservation et la communication, mais cela évite de poser les questions d'interprétation. Avec l'essor de la documentation numérique, certains auteurs entendent faire du concept de document une notion englobante. Or la documentation, discipline appliquée de la philologie, ne traite pas de la constitution des documents ni de leur lecture. En maintenant la distinction entre document, performance et œuvre, nous souhaitons toutefois souligner qu'ils relèvent de trois champs différents, objectivés par des disciplines diverses. Comment donc articuler ces disciplines pour réunifier ces niveaux de description et d'intelligibilité ? L'informatique n'a accès qu'aux documents. Or les unités propres aux performances sémiotiques ne correspondent pas nécessairement à des unités documentaires[5].

45 (ii) La performance est la *teneur* d'un document, son signifiant étant conventionnellement autonomisé de son support : dans les termes de la sémiotique de Hjelmslev, le support documentaire relève de la substance de l'expression et le signifiant de sa forme, la sémiotique étant définie comme science des relations et donc des formes ainsi comprises. L'autonomisation voire la séparation de la performance et du document doivent beaucoup à la pratique antique de la copie, puis à l'imprimerie, enfin à l'Internet, où la matérialité du support devient d'autant plus évasive qu'il n'est plus un objet mobilier comme furent le tableau, le rouleau ou le codex.

46 (iii) Appliquée aux performances sémiotiques, la notion d'*œuvre* dépend de domaines critiques qui s'attachent à l'évaluation et la description des inégalités qualitatives. L'élaboration particulière des œuvres procède d'un engagement pratique singulier, qu'il soit esthétique ou éthique. Elles se caractérisent par un appariement spécifique entre les plans du contenu et de l'expression, qui se traduit par une sémiosis unique.

47 *Rudiments épistémologiques* – Sur le plan épistémologique, il nous incombe de caractériser et d'individualiser les objets culturels de manière qu'ils deviennent lisibles et le demeurent. Il s'agit d'un processus progressif, mais sans fin, car aucune lecture scientifique n'épuisera un texte ; en revanche, on peut problématiser ses lectures, les rapporter à leurs conditions et les hiérarchiser.

48 Une première voie peut être dite *nomothétique* : elle résume la science à la formulation de lois – d'où, en sémiotique la tentation récurrente de la normativité, pour conformer le

réel à ses lois supposées. La voie inverse peut être dite *idiographique* : connaître, c'est alors caractériser la spécificité et la singularité des objets singuliers.

49 La force spécifique des sciences de la culture, c'est qu'elles peuvent – et doivent – articuler ces deux voies, dans un va-et-vient constant entre types et occurrences, règles et manifestations, instances et performances. Une règle est une norme qui paraît s'appliquer partout à une époque donnée. Quand bien même les tableaux de Rembrandt obéiraient aux mêmes normes de style, il reste à les différencier entre eux, et une telle tâche peut, au niveau d'analyse qui est le sien, incomber à la sémiotique de corpus. Alors que l'informatique est née pour une bonne part pour résoudre des problèmes de cryptographie (Turing), les premières analyses statistiques ont été utilisées pour résoudre des problèmes d'attribution de textes à leurs auteurs.

50 Étendant le champ d'investigation de la sémiotique générale, la sémiotique de corpus lui permet tout à la fois une reconception de son objet et de ses théories. D'une part, elle permet de construire une observation des normes. D'autre part, elle permet de concevoir la sémiosis des performances, en mettant en évidence des corrélations multiples entre plan du contenu et plan de l'expression. Elle périme enfin des conceptions traditionnelles, en s'opposant au modèle newtonien de la science, et permet à la sémiotique de s'intégrer pleinement aux sciences de la culture.

6. Quantité et qualité

51 Une sémiotique de corpus doit affronter des problèmes nouveaux. Comment passer de la quantité à la qualité, bref des méthodes quantitatives aux évaluations ? Des unités documentaires aux formes sémantiques ? Du document numérique à la performance ? Comment enfin articuler des critères locaux (portant sur les signes) à des critères globaux portant sur le genre de la performance et le corpus où elle prend son sens ? Bref, comment articuler traitements quantitatifs et descriptions qualitatives ?

52 *Mesure pour mesure* – La sémiotique de corpus s'appuie sur des méthodes quantitatives, pour l'essentiel statistiques. Mais comment concevoir cette mesure sans quelques précautions ?

(i) Ce qui est mesurable n'est pas forcément intéressant – bien que la quantification ait fini par devenir synonyme de respectabilité scientifique.

(ii) Ce qui est fréquent ne l'est pas forcément non plus ; notamment, les fréquences absolues sont inutilisables. Par exemple, dans un texte, les mots les plus fréquents sont des grammèmes qu'on retrouve dans tous les autres textes de la langue.

(iii) Les traits « de forme » n'ont pas de poids statistique déterminable, du moins dans l'état de l'art.

(iv) Les unités rares, de fréquence 1 (les *hapax*), ou absentes (les *nullax*) sont tout aussi intéressantes et souvent caractérisantes (par exemple, le mot *homme* reste absent des sites racistes).

(v) Les éléments les plus caractérisants sont des corrélations qui peuvent relier des éléments peu fréquents. Les événements sémiotiques (si par exemple l'on cherche à détecter les créations de concepts) sont des passages inédits, donc des *hapax* combinatoires.

(vi) Le qualitatif peut échapper à tout dénombrement : si des séquences figées de l'expression chaînes de caractères peuvent être dénombrées, il n'est pas certain que les

unités sémantiques soient assez discrètes et stables pour l'être de la même manière. De simples cooccurrences sans poids statistique peuvent être révélatrices.

53 *Qualifier les nouveaux observables* – La force heuristique de la sémiotique de corpus tient à ce qu'elle peut mettre en évidence voire « faire émerger » de nouveaux observables, notamment : (i) Des associations entre éléments qualitatif et quantitatif (par exemple, en corpus littéraire, les *hapax* sont associés aux pronoms de troisième personne). (ii) Des inégalités qualitatives, notamment dans la topographie des performances (comme les rafales ou les inégalités distributionnelles dans les sections du texte). (iii) Des associations entre unités sémantiques (thèmes) et unités expressives, même « de bas niveau » (par exemple, code couleur html).

Les corrélations constatées doivent pouvoir être interprétées pour accéder au rang de faits nouveaux. Un nouvel observable est un générateur d'hypothèses et éventuellement un précieux destructeur d'évidences.

54 *Dépasser la contradiction quantité/qualité* – En 2015, Microsoft a scanné 360 des 400 toiles attribuées à Rembrandt, puis a entraîné un système connexionniste d'apprentissage des régularités, qui a donné ce verdict : le Rembrandt type portraiture un homme d'âge mûr, avec barbiche et moustache, qui porte collerette blanche sur habit noir, se coiffe d'un large chapeau et regarde vers la droite. Dans ce tableau léché, Microsoft garde la prudence qui décèle les mauvais faussaires : tout a l'air vraisemblable, car tout reste prévisible. Or Rembrandt n'a pas créé seulement le rembranisme, car il a su s'écarter de ses propres stéréotypes, avant et mieux que ses suiveurs. Par exemple, le célèbre portrait de Oopjen Coppit, dite la Joconde du Nord, ne correspond en rien au type idéal inféré par l'intelligence artificielle et ne « coche » aucun de ses traits définitoires : il représente une femme en pied, au bas d'un escalier, tenant un éventail, regardant vers la gauche.

55 Aussi, quantité et qualité forment-ils une dualité et ne sont interprétables que l'une par l'autre. En effet, la quantité ne signifie rien par elle-même et l'on ne peut s'en tenir par exemple à des fréquences absolues, car d'un point de vue comparatif, seules les fréquences relatives sont interprétables. La qualité se discerne et se caractérise par un raisonnement comparatif entre quantités : par exemple, un mot absent, éminemment qualitatif et souvent révélateur, n'a qu'une fréquence zéro.

56 Pas plus que le fréquent n'est assimilable au quantitatif, le rare ne se confond avec le qualitatif. Il n'y a pas d'opposition entre quantitatif (positiviste) et qualitatif (élitiste), mais une complémentarité : ainsi, le résultat quantitatif peut confirmer l'hypothèse qualitative. Fréquente ou rare, toute donnée numérique, fût-elle un zéro, doit être rapportée à une donnée textuelle. Or une donnée textuelle, c'est ce qu'on se donne. Par construction, elle est le lieu d'interaction de quatre pôles : Contenu et Expression, Point de vue et Garantie (cf. l'auteur, 2011, ch. 2). Ordinairement, à partir d'une expression, on doit reconstituer et qualifier les trois autres pôles pour objectiver la donnée. Les « données » sémiotiques sont ainsi le résultat d'une interprétation. D'autant plus que les sorties logicielles sont souvent ininterprétables : par exemple, le résultat graphique d'une analyse factorielle n'obéit à aucune métrique simple, et il faut bien connaître le corpus pour pouvoir l'interpréter. Ce n'est donc pas l'instrumentation qui permet l'interprétation, mais l'inverse.

57 Bien entendu, la qualité l'emporte sur la quantité. Les meilleurs algorithmes ne peuvent apporter quelque secours que si l'on a défini les données initiales de manière critique. Par exemple, les capacités d'un système connexionniste sont étroitement liées à la qualité du

corpus d'apprentissage à partir duquel s'établissent les poids de ses connexions. Ainsi, le détour instrumental, loin d'égarer dans le quantitatif, part d'un jugement qualitatif pour permettre de l'éprouver et de l'enrichir en découvrant d'autres qualités associées mais restées jusqu'alors insoupçonnées.

7. Incidences en retour sur la théorie sémiotique

58 Si, faute de paramètres historico-culturels reproductibles, l'expérimentation au sens strict reste impossible dans les sciences sociales, les observables produits par la sémiotique de corpus sont bien des phénomènes nouveaux : dans cette mesure, cette sémiotique instrumentée jouit d'un potentiel heuristique considérable. L'informatique n'est pour elle qu'un instrument, non un modèle théorique, car la sémiotique appartient pleinement aux sciences de la culture. Un traitement reste évidemment un objectif technique et non un objet scientifique : confondre les deux, ce serait constituer une technoscience qui trouverait sa légitimité dans les outils. Les traitements automatiques doivent leur scientificité à la sémiotique dont ils constituent un secteur d'application.

59 Il faut donc passer du principe de plaisir théorique au principe de réalité descriptive. Un nouveau rapport à l'empirique change non seulement l'étendue mais la nature des faits et rend nécessaire l'innovation théorique. Il permet de produire de nouveaux faits, qui naissent pour ainsi dire de la rencontre entre de nouveaux modes d'observation et d'explication.

60 L'accès aux corpus conduit ainsi à modifier le rapport entre théorie et pratique, tant en amont du processus de recherche, dans la formulation des hypothèses, qu'en aval, dans la recherche de contre-exemples ou de variations. C'est aussi le moyen de sortir des apories théoriques suscitées par la philosophie du langage : par exemple, l'analyse de corpus reste le seul moyen éprouvé pour relativiser et réduire la polysémie comme de contrôler l'ambiguïté ; ou encore, pour déterminer les valeurs des formes grammaticales : par exemple, le futur n'a pas les mêmes valeurs dans le discours juridique que dans le roman.

61 Enfin, la sémiotique de corpus, dès lors qu'elle adopte un point de vue réflexif à l'égard de ses propres démarches, peut permettre de rompre avec l'objectivisme candide : elle ne pratique pas d'analyse automatique des *données*, dans la mesure où elles doivent d'abord être qualifiées comme données, puis interprétées après traitement.

62 Pour cela, il faut élaborer une théorie de la sémiosis, qui loin d'être une lointaine extension de la linguistique, y occupe un rôle central, non seulement parce que la performance sémiotique est l'unité minimale d'étude, mais parce que la sémiosis globale de la performance détermine la sémiosis des paliers inférieurs et permet de concevoir l'unité du contenu et de l'expression.

63 Dans cet agenda, la sémiotique de corpus met l'accent sur deux complémentarités générales : celle des niveaux ou plans de description (comme pour les langues la morphologie, la syntaxe, la sémantique) et celle des paliers d'organisation et de complexité (comme le mot, la phrase, le texte, l'intertexte).

64 On peut bien entendu formuler l'hypothèse que paliers et niveaux correspondent à des variations objectives de complexité voire de statut empirique. Il reste cependant à problématiser ces variations sans explorer dans l'abstrait les complémentarités entre niveaux et paliers. En effet, les applications qui font l'objet d'une demande sociale croissante requièrent la mise en évidence de ces complémentarités : détecter un type de

site ; faire de l'analyse thématique assistée ; faire de la diffusion ciblée en définissant des proximités entre documents, etc. La plupart des applications supposent aujourd'hui des tâches de caractérisation : au sein d'un corpus, il s'agit de singulariser les éléments pertinents pour l'application. Dès lors, la sémiotique renoue, par une voie nouvelle, avec la problématique de la description des singularités, qui est propre aux sciences de la culture ; la description de lois, qui fut longtemps jugée la condition nécessaire de la scientificité, se subordonne alors à l'étude systématique des usages effectifs.

65 *Le problème épistémologique de l'objectivation* – L'essor de la méthode expérimentale dans les sciences de la culture n'est possible qu'en éliminant le subjectivisme et l'objectivisme : l'objectivité est la fin du processus et non son début.

66 (i) Dérivée de Simondon, appuyée sur Saussure en linguistique et sur Leroi-Gourhan en paléontologie et anthropologie, la théorie de l'individuation récuse les principaux postulats de l'ontologie, comme la permanence et la séparabilité des formes et des substances.

L'individuation est un phénomène différentiel, car elle résulte de la création de différenciations dans un champ parcouru de tensions. Aussi peut-elle être décrite dans le cadre d'une sémiotique différentielle qui généralise l'expérience et les acquis de la sémantique différentielle.

67 (ii) Comme le sens est fait de différences, l'activité scientifique le donne à percevoir et à comprendre en ménageant des différenciations. En raison des solidarités d'échelle, c'est un principe général, des signes élémentaires au corpus. En fonction des degrés de complexité, ces différents niveaux peuvent se complexifier : des « amas » de différences (comme les thèmes) peuvent cependant être identifiés et contrastés, notamment par des méthodes quantitatives.

Un corpus d'élaboration est un champ de prises de forme, et à ce titre il est parcouru de tensions en cours de différenciation. L'originalité des objets culturels tient à ce que leur prise de forme dépend du champ immédiat de leur élaboration, mais aussi des champs médiats que sont les corpus distants, et cela à deux degrés : celui du corpus de référence, voire de l'archive ; celui des corpus d'interprétation, qui interviennent ultérieurement dans les phases postérieures de l'individuation.

68 (iii) Dans ce cadre épistémologique, l'élaboration de la vérité scientifique n'est pas un processus d'*accrétion*, comme une accumulation inductive de petites vérités (qui supposerait une compositionalité), mais un processus d'*érosion* : la connaissance progresse en cherchant à infirmer des hypothèses, et les vérités scientifiques restent des conjectures, tout en pouvant s'avérer infrangibles – ou relatives à telle ou telle échelle de grandeur. Il est d'autant plus nécessaire de trouver les moyens d'infirmer des hypothèses : le détour instrumental qu'appelle la sémiotique de corpus permet de l'assurer, pour autant qu'il soit exploité de manière critique.

BIBLIOGRAPHIE

Bourion, E. (2001). *L'aide à l'interprétation des textes électroniques* (Thèse de doctorat). Université de Nancy II. Ed. pdf http://www.texto-revue.net

Mayaffre, D. (2002). Les corpus réflexifs : entre architextualité et intertextualité. *Corpus*, *I*(1), 51-70.

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

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

Rastier, F. (2013). *Apprendre pour transmettre – L'éducation contre l'idéologie managériale*. Paris : PUF.

Rastier, F. (2018). *Faire sens. De la cognition à la culture*. Paris : Classiques Garnier.

Saussure, F. de (2001). *Écrits de linguistique générale* (éd. S. Bouquet et R. Engler). Paris : Gallimard.

NOTES

1. Nous nous appuyons ici sur l'auteur, 2011, ch. 2.
2. Un texte en effet ne peut pas être produit par un système, comme l'a montré l'échec de la grammaire générative appliquée à des systèmes de génération automatique de phrases et *a fortiori* de textes.
3. Cette tripartition due au philosophe Charles Morris et au logicien Rudolf Carnap relève du positivisme logique et reste un obstacle épistémologique majeur pour la sémiotique.
4. Voir l'auteur, 2001, ch. 3 et 4.
5. Par exemple, le morphème et la lexie, unités linguistiques, ne correspondent pas clairement à ces unités documentaires que sont les chaînes de caractères et l'on sait les multiples difficultés qui en découlent pour les traitements automatiques du langage.

RÉSUMÉS

Comme l'accès à la culture passe de plus en plus par les documents numériques, le système éducatif n'est pas en reste, car tout l'y engage, des habitudes prises par les élèves et les enseignants, aux recommandations ministérielles voire aux libéralités intéressées des grandes firmes comme Google. Par ailleurs, dans diverses disciplines, de la linguistique à la musicologie, de l'iconologie aux études de cinéma, de l'archéologie à l'anthropologie, des corpus d'étendue et de qualité croissante sont constitués et leur investigation est assistée par divers logiciels. C'est la base empirique de cette étude qui entend favoriser la convergence entre l'éducation et la recherche en sciences de la culture – sans pour autant se mêler de recherche didactique. Dans ces deux domaines toutefois, l'accès aux documents n'est qu'une condition de leur appropriation et

l'on doit se protéger des flux de « données » hétérogènes, non garanties et de ce fait ininterprétables.

As access to culture is more and more using digital documents, the education system is not left out, because everything involves it, from the habits taken by students and teachers, to ministerial recommendations and even interested big companies like Google. In addition, in various disciplines, from linguistics to musicology, from iconology to film studies, from archeology to anthropology, corpora of increasing quality are constituted and their investigation is assisted by various software. It is the empirical basis of this study that aims to promote convergence between education and research in the sciences of culture - without interfering with didactic research. In these two areas, however, access to documents is only a condition of their appropriation and we must protect ourselves from heterogeneous, unsecured and therefore uninterpretable "data" flows.

INDEX

Mots-clés : corpus, sémiotique, sciences de la culture, interprétation, documents numériques
Keywords : corpus, semiotics, cultural sciences, interpretation, digital documents

AUTEUR

FRANÇOIS RASTIER

Directeur de recherche (CNRS, INaLCO-ERTIM)