(EN)-JEUX DE CORPUS POUR LA RECHERCHE EN SHS. ÉNONCÉS, TEXTES ET DOCUMENTS

Huguette RIGOT
INRP, Paris X

SOMMAIRE

Résumé : *La notion de corpus est à caractériser, connotant des réalités et des objets textuels différents suivant les disciplines et les situations de recherche. Pour ce faire, les pratiques issues de trois traditions savantes sont utiles à analyser.*
La linguistique de corpus fait figure d'exemple dans les SHS. La réflexion de la linguistique de corpus a permis non seulement de spécifier cette notion, mais aussi de développer de nouveaux champs disciplinaires souvent, d'ailleurs articulés à d'autres disciplines comme l'histoire, l'analyse de discours en est certainement le meilleur exemple.
Mais s'il existe aujourd'hui un dynamique ensemble de réflexions portant sur la notion de corpus, certains secteurs de la recherche n'ont pas encore intégré les potentialités des notions de corpus et de numérique. Quand pourra-t-on parler d'une sociologie du corpus ? Cette question que se posent aujourd'hui chercheurs et institutions est particulièrement pertinente et y répondre est urgent. La constitution de corpus de données d'enquêtes quantitatives et surtout qualitatives correspond à plusieurs objectifs qui convergent tous vers un repositionnement épistémologique et réflexif des sciences de la culture.

Introduction

Le chrononyme « société de l'information » nous projette dans l'ère du numérique. Il transforme de manière radicale notre rapport à l'écrit et notre rapport à l'autre. Les changements qui, actuellement, nous affectent, ont une dimension à la fois sociale, cognitive et sémiotique.
Pourtant, assez paradoxalement, la plupart des pratiques de recherche en sciences humaines et sociales ne sont concernées qu'à la marge par l'évolution ou la révolution apportée par le numérique.
Aussi, il semble intéressant d'analyser en premier lieu quelles sont les possibilités ouvertes par les nouvelles technologies de l'information, notamment par la constitution de corpus numériques et/ou de bases de données textuelles regroupant des matériaux issus d'enquêtes qualitatives, puis dans un deuxième temps de confronter ces possibilités à l'existant, par l'analyse du traitement actuel appliqué à ces données d'enquête par leurs producteurs et enfin de considérer comment le positionnement épistémologique des sciences humaines et sociales peut évoluer grâce à la constitution de ces corpus numériques.
La lecture de différents rapports sur le statut des données d'enquêtes qualitatives permet, à partir de la reconnaissance ou de la non-reconnaissance de la valeur des enquêtes qualitatives par diverses communautés académiques, de constater que ce problème est abordé de deux manières, soit par le biais de la langue donc des corpus oraux et soit par celui du recueil des

données, focalisé sur la pratique d'entretien générant des corpus de données orales à la fois nombreux et volumineux.
Notre hypothèse de travail est que ces deux voies sont complémentaires. Pourtant, en prenant les disciplines qui les symbolisent le mieux, à savoir d'un côté la linguistique de corpus et de l'autre la sociologie, celles-ci travaillent encore peu de concert : les objectifs de connaissance poursuivis et surtout les modalités méthodologiques sont différents. La linguistique de corpus est une discipline d'observation des pratiques langagières, alors que la sociologie et toutes les sciences sociales et humaines, utilisant comme méthode de recueil de données l'entretien, l'immersion participante, etc., sont des disciplines d'interactions, possédant une forte valeur communicationnelle.
Ainsi, d'un côté l'observation, de l'autre les interactions ! Cette opposition prenant en compte des modes d'approche spécifique des réalisations langagières permet de comprendre d'une part pourquoi les données qualitatives sont à ce point « abandonnées », en déshérence et servent d'argument à la dévalorisation des résultats obtenus par les sciences humaines et sociales quand ceux-ci ne doivent rien aux statistiques et d'autre part comment se donner les moyens de modifier le statut épistémologique des sciences humaines et sociales en instaurant de nouvelles approches méthodologiques du traitement des paroles recueillies auprès de la société civile. Ces modes d'approche des réalisations langagières, en rendant compte d'un engagement différencié des chercheurs par rapport à ces données – observateurs des corpus oraux ou acteurs participant à des interactions avec des enquêtés – sont des facteurs explicatifs à la fois du statut méthodologique accordé aux corpus oraux et des traitements qui leur sont appliqués.

1. Les corpus numériques et / ou bases de données textuelles : pourquoi et comment les constituer ?

1.1. Le corpus : une notion désormais centrale

En évoluant de l'introspection au corpus[1], aujourd'hui, la linguistique se définit principalement comme une discipline d'observation des choix linguistiques effectués par des locuteurs dans des contextes réels. Par ce changement, quatre éléments possibles émergent des analyses linguistiques :

- le corpus remplace le texte, qui lui-même a été l'objet d'une sorte de révolution en devenant une unité d'analyse plus complexe que le mot et la phrase,
- les réalisations langagières sont produites par des acteurs réels, ordinaires et ou appartenant à des communautés spécifiques. La variation linguistique est devenue ainsi un objet d'analyse, tout en s'appuyant sur des faits de langue authentiques.
- les possibilités d'accéder à des réalisations langagières sont devenues infinies et bien évidemment, le web peut être aujourd'hui d'une part considéré comme un réservoir illimité à la fois par le statut, la variation des réalisations présentes et surtout par leur volume et d'autre part comme le moyen d'accéder à des bases de données textuelles et des corpus numériques constitués spécifiquement pour l'analyse linguistique, la base de données Frantext en est l'exemple le plus significatif.
- la comparaison des pratiques langagières est désormais devenue non seulement possible, mais un des fondements de cette nouvelle linguistique.

1.2. Qu'est-ce qu'un corpus ?

Si cette notion de corpus linguistique est devenue si centrale, elle nécessite d'être définie et caractérisée. François Rastier, sous forme métaphorique, parle de « sac de mots ou archives de textes[2] », évidemment pour proposer une caractérisation plus scientifique.
La notion de corpus renvoie traditionnellement à deux conceptions. La première est documentaire et ne retient que des variables globales ignorant les aspects textuel et structurel. Dans ce cas, le corpus est un réservoir d'exemples langagiers ou… une base de données textuelles.

[1] Pour s'orienter efficacement sur quelques textes fondamentaux et récents sur la linguistique de corpus, on peut citer : *La linguistique de corpus* sous la dir. de Geoffrey Williams, Presses universitaires de Rennes, 2005 et la revue *Corpus*, Numéro 1 « Corpus et recherches linguistiques », novembre 2002. Notamment les articles de Jean-Philippe Dalbera, « Le Corpus entre données, analyse et théorie » et Damon Mayaffre, « Les corpus réflexifs : entre architextualité et hypertextualité ».
[2] François Rastier, « Enjeux épistémologiques de la linguistique de corpus » dans *La Linguistique de corpus* sous la dir. de Geoffrey Williams, Presses universitaires de Rennes, 2005, p. 31.

La deuxième conception, plus liée à une tradition herméneutique, prend en compte les relations intertextuelles.
La définition proposée par F. Rastier fait du « corpus (…) un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications » [1].
Ainsi, tout regroupement de textes ne peut être considéré comme un corpus… il peut être simplement une base de données textuelles.
Un corpus suppose une préconception, c'est-à-dire une explicitation du choix des textes qui le composent et donc une sélection. Il implique la définition d'un objectif d'usage, donc d'un projet scientifique. Cette caractéristique est fondamentale, puisqu'elle permet de dire que le corpus n'est pas représentatif de la langue, mais qu'il est un **construit**[2] adéquat à un objet ou à une tâche qui détermine alors ses critères de représentativité et son degré d'homogénéité. Dans tous les cas, le corpus dépend du point de vue académique et aussi « personnel » du chercheur. Le corpus doit être à la fois construit et « aimé » par celui qui le rassemble et l'analyse.
S'il est un construit, le corpus est situé dans des pratiques qui « travaillent », documentent, catégorisent des rassemblements textuels, ainsi, quatre niveaux sont à distinguer : l'archive qui regroupe l'ensemble des documents accessibles, le corpus de référence à partir duquel des corpus d'études vont être délimités et enfin le sous corpus de travail variant selon les étapes de l'analyse.
Ainsi, la notion de corpus linguistique étant stabilisée et opératoire, de quelle façon peut-elle être utilisée pour organiser des rassemblements de textes en sciences humaines et sociales ? La réflexion, qui a conduit à constituer la linguistique de corpus peut-elle aider à faire comprendre l'importance, du point de vue de la linguistique et plus généralement du point de vue des sciences humaines et sociales, du rassemblement, donc de la préservation en vue de la communication des matériaux langagiers produits par les enquêtes qualitatives ?

2. Préserver pour communiquer : une nouvelle approche des matériaux langagiers en sciences humaines et sociales

Il s'agit bien évidemment des données qualitatives, entretiens enregistrés, notes écrites lors d'observation participante, etc.
De ce fait, il faut écarter les données quantitatives. Pourtant, en France leur traitement n'est peut-être pas sans rapport avec celui dévolu aux données qualitatives.
Leur situation a été récemment et peut-être momentanément réglée à partir du Rapport[3] de Roxane Silberman du Lasmas - CNRS qui constatait que les données publiques produites avec des financements publics n'étaient pas accessibles directement aux chercheurs individuels et aux laboratoires, les grands producteurs de données quantitatives comme l'INSEE faisant payer leur utilisation. Ainsi est né le centre Quetelet[4] dont la fonction est d'archiver les données des différents producteurs de la Statistique publique à des fins de diffusion et de réutilisation pour les chercheurs. Cette utilisation fait l'objet d'une réglementation. Cet accès est limité aux fonctions de recherche ou d'enseignement, pour écarter les réutilisations purement commerciales et il se fait selon une réglementation et à partir d'un engagement formel et écrit des utilisateurs. Le problème de la préservation pour communication des données quantitatives semble réglé.
On peut raisonnablement penser que cette nouvelle modalité d'accès permet de dynamiser les recherches. La plupart des laboratoires, avant la création du Centre Quetelet, se contentaient de citer des données sous forme de schémas, tableaux, etc. repris de publications précédentes. Ainsi, le lecteur averti lisait plusieurs fois les mêmes informations sans avoir vraiment l'impression d'une réutilisation critique. Mais pour critiquer les sources… faut-il encore y avoir accès.

---

[1] *Op. cit.* p. 32
[2] Marie-Paule Jacques, « Pourquoi une linguistique de corpus ? » dans *La Linguistique de corpus* sous la dir. de Geoffroy Williams, Presses universitaires de Rennes, 2005, p. 26.
[3] Consultable à l'adresse http://www.ladocumentationfrancaise.fr/rapports-ublics/004000935/0000.htm (consulté 28 juin2006)
[4] Consultable à http://www.centre.quetelet.cnrs.fr/

2.1. La variété et le statut des matériaux langagiers

Que ce soit dans les rapports Français ou étrangers, quand on évoque ces corpus numériques de données qualitatives, il est quelquefois difficile de savoir de quel type de données on parle. Nous sommes dons assez loin de la caractérisation précise de la notion de corpus telle qu'elle est présente dans la linguistique. De façon assez large, il peut s'agir de trois types de données :

a) des revues en lignes et des archives ouvertes qui mettent à la disposition de lecteurs sur le net des résultats d'enquête publiés ou finalisés auxquels on accède par abonnement ou librement.

b) le plus souvent, il s'agit de données d'enquête non publiées : la « littérature grise » et qui pose là un véritable problème de préservation et de communication tant auprès des chercheurs qu'auprès d'un public plus élargi. À cet égard, le *Rapport Canadien de mai 2001 sur l'évaluation des besoins sur la Consultation nationale sur les archives de résultats de recherche*[1] est tout à fait révélateur de l'ambiguïté ou de la difficulté à définir et à caractériser l'objet de la conservation : littérature grise uniquement ou matériaux d'enquêtes qualitatives ?

c) d'autres rapports indiquent clairement que ce sont les matériaux des enquêtes qualitatives qu'il faut préserver en vue d'une communication. Même si le mode d'accessibilité souhaité est le net, l'état des données est à ce point préoccupant que le simple repérage d'ensembles documentaires, consultables dans des lieux publics, comme les dépôts d'archives, correspondrait déjà à une révolution dans le traitement des matériaux qualitatifs.

2.2. La déshérence des données, une raison de la dévalorisation des enquêtes qualitatives, le cas français

En avril 2003, Françoise Cribier, avec la collaboration d'Elise Feller, a rédigé et présenté au Ministère délégué à la Recherche et aux nouvelles technologies un rapport portant sur la conservation des données qualitatives des sciences sociales en France[2]. La rédaction de ce rapport se situe dans un contexte d'interrogation sur le statut et surtout le devenir des données issues des enquêtes qualitatives, c'est-à-dire des données orales. En l'espace de moins de dix ans plusieurs rapports ont été commandés par différentes institutions[3] et rédigés par des spécialistes se positionnant différemment, selon les commanditaires. C'est d'abord sous l'angle des archives sonores de diction – une place importante a été faite aux archives radiophoniques et télévisuelles – puis sous l'angle de l'archivage d'entretiens faits à l'occasion d'enquêtes qualitatives et enfin, sous l'angle de la conservation des fonds sonores que cette problématique a été abordée. Ainsi, trois préoccupations s'articulent ou s'entremêlent : les traces laissées par les pratiques langagières, les données d'enquêtes qualitatives et la préservation des documents sonores. Indiquer ce contexte n'est pas inutile pour comprendre comment une synergie se met en place ayant pour objectif de traiter, pour ce qui nous concerne ici, des données d'enquêtes qualitatives. Un entretien, par exemple, effectué dans le cadre d'une enquête, ressort bien à la fois de l'usage de la langue, de l'information scientifique qu'il a contribué à produire et enfin du support sur lequel il a été enregistré. Ce support est double. Il peut s'agir du support matériel, c'est-à-dire de bandes

---

[1] Consultable à http://www.sshrc.ca/web/about/publications/da_phase1_f.pdf (consultée 28 juin 2006)

[2] Françoise Cribier, *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile ».* Rapport présenté au Ministère délégué à la Recherche et aux nouvelles technologies en avril 2003.
http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf, consultée le 15 juin 2006.

[3] Les principaux rapports sont :
- le rapport remis au Conseil économique et social en janvier 2001 par Georgette Elgey avec la collaboration d'Annette Wieworka portant sur l'ensemble des archives sonores de diction
- le rapport, sous la direction de Claude Dubar et du Laboratoire Printemps en 2001 pour le secteur SHS du CNRS, s'intéressant surtout aux entretiens d'enquête
- le rapport rédigé par Marie-France Calas en 2001 pour le Ministère de la culture et de la communication aborde les problèmes de politiques de conservation et de valorisation menées en France pour l'ensemble des fonds sonores.

Il convient aussi de citer le *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*, sous la direction d'Olivier Baude et commandé par la Délégation générale à la langue française et aux langues de France. Ce guide vient d'être édité sous le nom de *Guide corpus oraux* par les éditions du CNRS.

audio et/ou audiovisuelles, mais aussi du support des transcriptions transformant les entretiens d'énoncés oraux en textes écrits qui, eux aussi, sont liées à des types de supports différents. Une transcription d'entretien, sans aborder tout de suite le problème du codage ou du « nettoyage » des données, suppose sa transformation en texte écrit sur du papier ou en fichier informatique.
Autre remarque, les disciplines n'ont pas toutes le même rapport à la parole de l'interviewé, psychologues et ethnologues savent mieux conserver la parole d'autrui. Leur formation les conduit d'une part à accorder une plus grande importance à ce type de matériau qui est souvent issu de situations singulières – l'entretien clinique, confidentiel pour le psychologue, l'observation, la participation, les rencontres exotiques pour l'ethnologue – et d'autre part à savoir comment matériellement les documenter pour les retravailler plus tard ou collectivement. Peu de sociologues ont ce souci de préservation et de partage.
En France, la situation des données d'enquêtes qualitatives est très préoccupante. Comment spécifier ce « contexte alarmant de l'archivage des sources orales »[1] ? Pour l'heure, aucune institution, aucune communauté – sauf peut-être celle des ethnologues – très peu de laboratoires et enfin de très rares chercheurs ont pour préoccupation de sauvegarder leur données d'enquête.
Au cœur du problème soulevé par la préservation et la sauvegarde de l'ensemble des données d'enquête, se situent la comparaison, la cumulativité des résultats, c'est-à-dire l'inscription des recherches dans une articulation de temporalités, celle de la recherche elle-même, celle des effets produits et enfin celle du regard distancié par le temps, les changements sociaux et les paradigmes scientifiques, qui sont directement en cause. Dans la majeure partie des cas, les résultats d'enquêtes, même quand ils sont publiés, sont accessibles au grand public et au public de chercheurs sous un format particulier, c'est-à-dire un nombre de caractères, de pages, un appareil informationnel plus ou moins développé, imposé par un éditeur à partir de contraintes commerciales et non scientifiques.
Données perdues, égarées, rangées dans les placards, détruites par manque de place ou à la suite de déménagements, donnent une image de la recherche et de l'importance de la méthodologie tout à fait contraire à celle qui prévaut dans les ouvrages de méthodologie d'enquête. De plus, le temps passé au recueil, au traitement, à la documentation de ces données est bien plus important que le temps passé à la rédaction des résultats de recherches. Pourtant ce sont ceux-ci qui restent, s'ils sont édités, qui servent à produire des interprétations, ce sont eux qui témoignent de la science-en-train-de-se faire, ce sont eux qui finissent par représenter la science faite[2]. Ce sont ces résultats de recherches qui servent à évaluer et la qualité de l'enquête et la qualité des chercheurs. À ce propos, il faut remarquer que, particulièrement dans le rapport rédigé par Françoise Cribier, la qualité de l'écriture des résultats d'enquêtes a orienté l'auteur vers certains chercheurs à interviewer, établissant ainsi une hypothèse implicite, peut-être à vérifier, à savoir que de bons travaux de recherche, des résultats bien rédigés et publiés sont sous-tendus par de « bonnes pratiques » de recueil et de traitement des données.
Force est de constater une fracture entre ce qui est dit, écrit dans les manuels de recherche à l'usage des jeunes chercheurs, – le terrain et les entretiens sont des actes quasi initiatiques – et ce que les chercheurs font réellement de leurs données, une fois les résultats écrits et publiés, c'est-à-dire un abandon, une mise au placard, un rejet. Mais, peut-il en être autrement en l'absence d'une formation à la sauvegarde des données, en l'absence de centres d'archivage et peut-être surtout en l'absence d'une conscience que ces données appartiennent aux chercheurs, aux enquêtés et aux institutions commanditaires, en l'absence d'une conscience de la valeur patrimoniale des paroles recueillies auprès de la société civile.

2.3. L'engagement du chercheur dans ses données

Si la situation d'un point de vue général est particulièrement alarmante, certaines communautés scientifiques, comme les ethnologues, et certains professionnels – archivistes, documentalistes – certainement sous la double impulsion donnée d'une part par la linguistique de corpus et d'autre part par le développement des nouvelles technologies, permettant la numérisation de grandes masses de documents et surtout leur mise à disposition, ont décidé de réagir et de mettre à profit leur savoir-faire professionnel et scientifique pour stopper la perte systématique des données qualitatives, et surtout pour renouveler en profondeur le travail scientifique.

[1] Rapport de Claude Dubar (ce rapport n'ayant jamais fait l'objet d'une large communication, n'est pas accessible).
[2] En référence aux travaux de sociologie des sciences, notamment ceux de Bruno Latour.

Le désintérêt pour ces données qualitatives porte témoignage d'une double conception de ce qu'est faire de la recherche en sciences humaines et sociales.
En premier lieu, il faut s'interroger sur la raison de la déshérence de ces données et comprendre que cette situation a parti lié avec le statut épistémologique des sciences humaines et sociales.
Les matériaux qualitatifs sont considérés comme peu dignes de confiance, ils témoignent d'un certain type de vérité, celle rapportée par les enquêtés, mais qui n'est pas celle des scientifiques. Travailler à partir de tels matériaux, c'est s'exposer à un certain nombre de biais scientifiques, quasiment tous répertoriés. D'abord ce sont des sources provoquées donc biaisées du fait des circonstances de leur élaboration. Ensuite, ils sont le produit de ce que Pierre Bourdieu appelait l'illusion biographique[1]. Il distinguait ainsi trois définitions de l'illusion : l'illusion téléologique surestimant l'intentionnalité car recomposant après coup des événements rassemblés pour atteindre un objectif, l'illusion de rester soi-même qui permet à l'individu de récupérer son unité à travers la complexité des situations vécues, et enfin l'illusion de personnalité permettant à l'individu de se sentir différent des autres.
En réponse à ces objections qu'elle considère comme étant réelles, Françoise Cribier, préfère admettre que ces trois grandes illusions correspondent aux réalités complexes de nos sociétés.
De plus, l'utilisation de ces matériaux qualitatifs demande à la fois une documentation sérieuse et une critique vigilante. Mais n'est-ce pas le travail même des scientifiques que de vérifier et de critiquer leurs sources ? Dans les recherches où ce travail n'a pas été mené, l'utilisation des matériaux qualitatifs est différente : au lieu d'étayer les éléments théoriques, ils servent d'illustration ou de confirmation aux résultats obtenus par d'autres sources. Ils sont donc détournés de ce qui fait leur spécificité et leur valeur : appréhender « les réalités » des acteurs ordinaires de la société civile.
En deuxième lieu et raison suprême à notre sens, la personne, la parole du chercheur tout comme celle de l'interviewé sont dans les matériaux qualitatifs. La situation d'interaction créée lors d'un entretien révèle autant sur l'un que sur l'autre et peut-être encore plus sur le chercheur lui-même, surtout quand il a « raté » son entretien, quand il n'a pas su entendre et lire l'importance de ce qui lui était transmis, quand il a surexploité ses données pour confirmer son hypothèse. *Les données, ça ne se partage pas, d'ailleurs ça n'aurait aucun sens,* voilà le credo de la plupart des chercheurs *et puis si ça tombait dans des mains peu amicales, des chercheurs concurrents pourraient « repomper » sans citer leurs sources, c'est ce qui est d'ailleurs arrivé à Bourdieu dans La Misère du monde, des chercheurs qui pourraient mal interpréter et ainsi alimenter des querelles stériles et puis qui va conserver tout cela, quelles institutions peuvent garantir la préservation, la communication et surtout la pérennité du travail considéré à juste titre comme bien plus important à cause de la documentation accompagnant les matériaux ?* Toutes ces questions sont posées dans les rapports. Elles témoignent de la même méfiance, de la même préoccupation : communiquer ce qui est personnel, sans l'avoir mis à distance par exemple par la mise en place de protocole de documentation, est difficilement acceptable. Les chercheurs préfèrent abandonner leurs données plutôt que de les partager, de les communiquer. Ils préfèrent les abandonner plutôt que de devoir les retravailler pour les transformer en corpus numérique.

2.4. Les modalités de constitution des corpus numériques

Je n'en évoquerai que les grandes lignes.
Premier principe : il est moins coûteux de prévoir et de préparer la constitution de ces corpus avant et pendant l'enquête que par la suite. Documenter une enquête est un travail habituel, mais écrire et catégoriser donc normaliser cette documentation est autre chose. Seuls les chercheurs ayant appris à le faire dans leur formation et ayant l'habitude de travailler collectivement et donc de partager leurs données savent le faire. Documenter, cela consiste entre autre, à spécifier le contexte général de l'étude et chacune des situations d'enquête créée, comme les entretiens. Cela consiste aussi à établir des connexions avec d'autres situations, à garder la trace systématique des impressions, des événements particuliers qui peuvent sur le moment ou après permettre d'approfondir et de réorienter certaines interprétations.
Deuxième principe : respecter les dispositifs juridiques, donc faire signer un accord aux interviewés pour être questionnés et pour que leur parole soit communiquée en partie ou intégralement. Enfin, garantir l'anonymat.

[1] Pierre Bourdieu, « L'illusion biographique » dans les *Actes de la recherche en sciences sociales*, 1986

Troisième point, le plus difficile à mettre en place, c'est-à-dire le traitement des données : les problèmes techniques relatifs à la prise de son et à l'enregistrement de terrain, puis les problèmes relatifs aux transferts sur des supports durables et enfin, les problèmes plus « intellectuels » de la transcription, du codage, du nettoyage des données et enfin de l'analyse de la qualité des données qui engagent directement le quatrième point…
Quatrième point : Que sélectionner ?
Cinquième point : Comment constituer des bases de données textuelles… Quelle plate-forme logicielle ? Quel financement ? Quelle garantie institutionnelle ?
Sixième et dernier point qui relève à la fois de la maîtrise et des choix des chercheurs et des formats de données numériques : quels logiciels d'analyse des données ?

3. L'impact sur le statut épistémologique des sciences humaines et sociales

Le regard porté sur les méthodologies qualitatives est souvent ambivalent : d'un côté, une reconnaissance de richesse et de l'autre une méfiance (peu de généralisation, peu ou pas de contrôle). Pourtant, nombre de chercheurs, sociologues (cf. *Enquête de terrain*, dir. Par D. Céfaï[1]) spécialistes des SIC (cf. *Dictionnaire des recherches qualitatives en sciences humaines*, dir. A. Mucchielli[2]) ont travaillé à élaborer des recueils de méthodologies qualitatives pour chercher à comprendre les phénomènes sociaux à travers les réalisations langagières des individus.
Notre propos est de considérer comment, dans cette filiation d'élaborations méthodologiques, la constitution de corpus de données d'enquête et leur mise en accès sur le réseau devraient permettre d'ouvrir les méthodes qualitatives à de nouvelles perspectives, d'une part, pour leur donner une place reconnue et acceptée par les acteurs des sociétés civiles et politique et par ceux de la communauté scientifique et d'autre part pour améliorer leurs « performances » épistémologiques par la pratique de la cumulativité et de la comparabilité des données, permettant ainsi de produire des méta-analyses.
Ainsi, on peut faire l'hypothèse que la déshérence des matériaux qualitatifs a contribué à disqualifier les recherches qualitatives qui sont, de par leur objectif et de par leur méthodologie, des modalités d'approche du social, uniques et spécifiques. La maltraitance des données d'enquêtes qualitatives portant à la fois sur l'impréparation du traitement et sur leur abandon, une fois publiés les résultats, a un impact direct sur le développement des sciences humaines et sociales. Faire ce constat, c'est donc se demander pourquoi changer, pourquoi transformer les pratiques, pourquoi conserver ces données.
La **première raison est d'ordre patrimonial** : les paroles recueillies dans des contextes particuliers d'enquête ne pourront jamais l'être de nouveau. C'est un matériau qui périt avec celui qui l'exprime et qui disparaît une fois l'interaction entre l'enquêteur et l'enquêté interrompue.
La **deuxième raison est la réutilisation de ces matériaux**. La lecture par un autre chercheur au moment de l'enquête ou plus tard, c'est-à-dire un autre regard social et culturel porté sur ce qui a été dit et observé, peut engendrer une autre interprétation. De plus, des comparaisons avec des travaux de la même époque ou des travaux d'époques différentes peuvent être établies. C'est la cumulativité des matériaux et des résultats de recherche et leur insertion dans une temporalité qui permettent une évolution significative de la recherche en sciences humaines et sociales.
La **troisième raison est d'ordre méthodologique**. Ce retour sur les matériaux implique aussi un regard rétrospectif sur les méthodes utilisées par les chercheurs et sur les conditions de production de leurs travaux. Une histoire des disciplines ou des champs disciplinaires serait ainsi possible. De plus, des chercheurs d'autres disciplines, en accédant à ces données « nouvelles » pour eux, pourraient ainsi développer la pluridisciplinarité et montrer la capacité des sciences sociales à explorer ensemble des objets communs, à fédérer les savoir-faire, à capitaliser les résultats, pour les revisiter, les réinterpréter et démultiplier les interprétations.
La **quatrième raison** considère que le regard porté sur l'acteur social n'est pas uniquement de l'ordre de l'objet d'étude : il se situe aussi à l'intérieur du processus de recherche et il est une figure de **l'engagement social et scientifique** que se doit d'avoir le chercheur.
La cinquième raison porte sur la complémentarité entre qualitatifs et quantitatifs. S'il y a peu de temps, les milieux académiques ont réagi pour assurer une conservation et une accessibilité aux données quantitatives, le même travail doit être accompli pour les données qualitatives, car de nombreuses recherches mêlent les deux approches.

[1] Daniel Céfaï , *L'Enquête de terrai*n, textes réunis, présentés et commentés, La Découverte, 2003.
[2] *Dictionnaire des méthodes qualitatives en sciences humaines* sous la dir. d'Alex Mucchielli, A. Colin, 2004.

Aborder le traitement des matériaux qualitatifs, c'est toucher réellement la méthodologie des enquêtes qualitatives. Transformer les pratiques d'enquêtes, la façon d'évaluer leurs procédures et leurs résultats, devient possible.
Reconnaître l'aspect communicationnel[1] de la recherche qualitative, c'est intégrer l'engagement du chercheur dans le processus d'enquête, accepter le partage des données qualitatives et démultiplier l'interprétation. Ainsi, les sciences humaines et sociales peuvent sortir de leur situation de disciplines dominées.
La linguistique de corpus pointe la nécessité de rassembler des matériaux langagiers réels, de les documenter pour décrire leurs conditions de production et rendre possible leur communication et d'effectuer des traitements interprétatifs. Elle met à disposition des chercheurs des sciences de la culture une méthodologie, une réflexion conceptuelle et surtout des modes opératoires rendant compte de la valeur spécifique de ces données. Elle inscrit celles-ci simultanément dans une temporalité, un cadre de pensée et des modalités de traitement assistés ou non par ordinateur. La notion de corpus absorbée, réappropriée et adaptée par les chercheurs peut sortir les recherches qualitatives de leur ghetto épistémologique.

BIBLIOGRAPHIE

CEFAÏ, D. 2003. *L'Enquête de terrain,* Paris, La Découverte.
Corpus. Numéro 1 « Corpus et recherches linguistiques », novembre 2002.
CRIBIER, F. 2003. *Projet de conservation des données qualitatives des sciences sociales en France auprès de la « société civile »*. En ligne sur le site du Lasmas (CNRS-EHESS).
Dictionnaire des méthodes qualitatives en sciences humaines, 2004. Paris, A. Colin.
RIGOT, H. 2005. Crises communicationnelles dans le processus de recherche en sciences humaines et sociales, in M. Gabay, *Communiquer dans un monde en crise*, Paris, L'Harmattan.

[1] Huguette Rigot, « Crises communicationnelles dans le processus de recherche en sciences humaines et sociales » dans Michel Gabay, *Communiquer dans un monde en crise*, Paris, L'Harmattan, 2005.