Dr. Tabet Aoul Zoulikha
MCB USTO MB

2. Notions de corpus et la base de données

Lorsque Greimas parle du lecteur « happé » par l'objet de son activité, on ne peut s'empêcher à la transformation fondamentale de la relation entre sujet et objet qui établit un nouvel « état des choses », c'est dans cette optique que s'inscrit cette contribution

Plan

1. les types de corpus

Même s'il s'agit d'une rubrique redondante dans les réflexions des membres de l'équipe du présent projet, l'importance de cette notion, dans la constitution de la base de données, est telle qu'il parait impératif de devoir l'aborder sous tous les angles possibles. Dans cette optique, nous revisitons les notions de corpus avec l'éclairage de la linguistique et de l'analyse du discours en particulier.

Corpus

Dans le vocabulaire des sciences, corpus désigne un recueil large, et quelquefois exhaustif, de documents ou de données : corpus de textes juridiques, corpus des inscriptions calligraphiques, corpus des discours oraux...

2. En linguistique :

Dans les sciences humaines et sociales tout particulièrement, corpus désigne les données servant de base à la description et à l'analyse d'un phénomène. En ce sens, la question de la constitution du corpus est déterminante pour la recherche puisqu'il s'agit, à partir d'un ensemble clos et partiel de données, d'analyser un phénomène plus vaste que cet échantillon.

Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites. On peut donc se questionner sur les méthodologies de constitution des corpus en termes de représentativité quantitative et qualitative par rapport aux phénomènes à décrire et à analyser : un corpus doit fonder des analyses objectivables et sa représentativité peut dépendre de sa taille. Il est cependant, dans la pratique, très délicat de définir avec précision la taille du corpus qui garantirait sa représentativité. De plus, la taille d'un corpus dépend aussi, pratiquement, de la possibilité de recueillir des données les stocker et de les préparer pour le traitement, ainsi que de les traiter. Les corpus peuvent être constitués par des données orales, écrites, audiovisuelles…

3. En analyse du discours :

La question semble se poser dans des termes voisins, compliqués cependant par le fait qu'il s'agit de décrire des phénomènes discursifs qui se déploient sur des surfaces textuelles importantes. On privilégie donc les corpus de grande taille (ensembles de textes, le plus souvent), lui sont traités manuellement, mais aussi par des procédures informatiques de traitement automatique, qui ont d'ailleurs présidé à

l'émergence du domaine (Pêcheux 1969). On pense alors la question primordiale de la représentativité statistique de données inédites, lesquelles pourraient être identifiées et recherchées à partir de la définition explicite du problème à traiter, par exemple à propos des textes médiatiques, quelle quantité d'exemples peut être considérée comme significative ?
Cependant, en analyse du discours comme dans d'autres sciences sociales, c'est souvent le corpus qui, en fait, définit l'objet de recherche qui ne lui préexiste pas. Ou plutôt, c'est le point de vue qui instruit un corpus.

Le mode de constitution du corpus n'est donc pas, en analyse du discours, un simple geste technique répondant aux exigences ordinaires de l'épistémologie des sciences sociales : il est problématique en ce qu'il met en jeu la conception même de la discursivité, de sa relation avec les institutions et du rôle de l'analyse du discours.
La relative nouveauté de la discipline de l'analyse du discours, rapportée à la masse des textes encore à décrire, le caractère souvent irréductible des points de vue fondateurs adoptés, invitent à la prudence et au débat, quand il s'agit de généraliser des résultats ou d'en proposer des explications qui, par définition, ne sauraient être internes à l'analyse du discours, mais qui convoquent l'ensemble de la société.

4. En lexicométrie :
La question du corpus prend un tour particulier car au moins le temps d'une expérience le corpus prendra une forme fermée, on ne peut compter en effet que sur un ensemble stabilisé. Le corpus sera la base des comparaisons permettant la contrastivité des invariants constitutifs. Le temps d'une expérience, la variable d'étude dépendra des hypothèses mises au départ dans la constitution du corpus.
Basés sur des calculs de probabilité, de statistique et d'hypergéométrie, les corpus en lexicométrie fourniront en sortie machine des matériaux divers susceptibles d'être indexés, triés, sélectionnés, hiérarchisés, lemmatisés, articulés les uns aux autres…
Matériaux qui constitueront autant de clés de lecture et d'analyse que le l'esprit critique et la compétence du chercheur auront à déployer.

5. Corpus journalistique :
Cette rubrique est traitée par les membres du projet travaillant sur le discours médiatique

6. Corpus d'archive :
Pour les historiens et les linguistes qui travaillent ensemble autour de l'analyse du discours, le corpus est l'ensemble des énoncés s'organisant en série qu'ils vont soumettre aux procédures rigoureuses de la linguistique. Le corpus peut être homogénéisé en fonction de l'appartenance idéologique des sujets ou de la conjoncture historique.

Cette présentation de corpus s'inscrit dans une réflexion commune, distincte d'un simple partage ou d'une addition de compétences, en mettant l'accent sur la problématisation du texte, du discours et du sens comme objet des sciences humaines et sociales afin d'arriver à poser l'analyse des formes langagières au centre des interprétations, des hypothèses de lecture.

Bibliographie

Garric N. & Léglise I. (2005). « La place du corpus, de l'analyste, du logiciel », in G. Williams (éd.), La linguistique de corpus. Rennes : Presses Universitaires de Rennes,2005, pp. 101-114.

Rastier F., « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (éd.), La linguistique de corpus. Rennes : Presses Universitaires de Rennes, 2005

Reinert M.,. « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », Langage et société 60,1993