Master « Linguistique »

Kheira Mérine

**Le corpus ; objet, concept et théorie**

**Section 1**

1. **Compétence :** constitution et classification d’un corpus
2. **Définition**

* Selon le point de vue des dictionnaires
* D’après le Larousse en ligne[[1]](#footnote-2), le corpus est un « ensemble fini d'énoncés écrits ou enregistrés, constitué en vue de leur analyse linguistique ».

*-*Pour le Robert, le corpus est un « recueil de pièces » (cité par Dalbéra)

*-* Wikipédia le définit comme étant « Un ensemble de documents, artistiques ou non regroupés dans une optique précise ».[[2]](#footnote-3)

* Selon le point de vue des linguistes.

Dalbéra (op.cit) distingue la définition d’ordre général de la définition propre aux sciences du langage.

Pour lui, d’une manière générale, « un corpus est un recueil de pièces ou de documents qui concernent une même matière, discipline ou doctrine » (Dalbéra, 2002).

Alors que dans les sciences du langage, cette notion renvoie à « un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique. »

1. **Etymologie et bref aperçu historique du concept.**

D’après le Trésor de la langue française, la première attestation lexicographique est de 1845 avec l’entrée suivante : **corpus juris**. Mots latins par lesquels on désigne le recueil des lois romaines tel qu'il a été fait sous le règne et d'après les ordres de l'empereur Justinien).(TLF [en ligne][[3]](#footnote-4)).

Le mot ***corpus*** y est également considéré comme étant un emprunt du latin mais aussi de l’Allemand ***Korpu*s** qui renvoie à « un ensemble de données d'un certain type (notamment en lettres) réunies en vue de leur étude scientifique (terme technique de la recherche scientifique) » (attesté depuis 1787) (ibid)

Ces différentes définitions se rejoignent en montrant que le terme corpus renvoie à un ensemble d’éléments écrits plus ou moins longs partageant des points communs qui peuvent être, d’ordre temporel, d’ordre thématique ou de spécialité. Ainsi, d’après Dalbéra (op.cit.), dès le XIXème siècle, le mot Corpus devient alors un mot français à part entière (ayant perdu sa forme latine), et l'on parle de *Corpus des Troubadours* ou de *Corpus* des poètes de la Renaissance. » (Dalbéra, op.cit.).

Dans cette acception, le corpus peut servir à des chercheurs appartenant à différents domaines (l’historien, le philologue, le juriste, l’anthropologue, le littéraire, etc.), même si leurs démarches varient d’un domaine à l’autre avec des objectifs différents.

1. Types de corpus et classification :

Tout corpus présente des critères qui lui permettent d’appartenir à tel ou tel type pour être classé dans telle ou telle catégorie. C’est ainsi qu’on peut distinguer deux sortes de corpus : ceux constitués de textes et ceux constitués de phonèmes, de mots, d’expressions ou de phrases.

Chaque catégorie peut présenter des sous-catégories à distinguer entre elles. Exemple :

* Ce qui relève de l’écrit et ce qui relève de l’oral.
* le corpus existant et le corpus fabriqué.
* Le corpus brut et les corpus non brut (arrangé selon le besoin)..

1. Le corpus en linguistique.

Quand le linguiste use de corpus c’est toujours pour en étudier un sous-ensemble de faits qui peut être de nature à rendre compte de l’ensemble. Pour cela, souvent on a recours à des échantillons qui ne devraient pas être choisis aléatoirement.

Concernant le corpus oral, on devrait distinguer le corpus servant à une étude d’interactions (interactions radiophoniques, interactions de type commercial ou professionnel, interactions entre des locuteurs réunis pour le même motif, etc.) d’un corpus où la communication est unidirectionnelle (entretien avec un spécialiste, un non spécialiste, un discours politique ou scientifique, une conférence, un plaidoyer, etc.)

Les enregistrements peuvent se faire en fonction des lieux choisis selon l’objectif assigné à la recherche (par exemple dans un magasin pour: étudier les interactions commerciales, dans un café ou devant son poste de radio pour étudier les tours de parole, les salutations dans les interactions radiophoniques, dans une salle de conférence ou dans un tribunal (plaidoyer),etc.). Il est à préciser que le corpus enregistré doit être soumis à une transcription selon des normes choisies.

Pour le corpus écrit, il peut relever du donné (généralement des textes complets pris par ensemble, ex trilogie, ou par unité, un texte à la fois), comme du fabriqué (texte tronqué, énoncés pour l’étude d’une structure ou mots et expressions pour une étude lexicologique).

Dans le corpus oral comme écrit, la construction à l’aide de données sélectionnées traduisent la préoccupation du chercheur et éclairent sur sa nature.

1. http://www.larousse.fr/dictionnaires/francais/corpus/19410#0uRELqcPglKysWDV.99

   [↑](#footnote-ref-2)
2. <https://fr.wikipedia.org/wiki/Corpus>, page consultée le 30/9/2016 [↑](#footnote-ref-3)
3. TLF [en ligne], page consultée le 1/10/2016 [↑](#footnote-ref-4)