LA LINGUISTIQUE DE CORPUS

Master 1 Sciences du langage : S2, 2025-2026

Dr AYAD Abderrahmane

Département de français

Université de Jijel

Mél : **abderrahmane.ayad@univ-jijel.dz**

Canal : **https://t.me/Cerclelinguistique**

◙ **Plusieurs ouvrages et articles en PDF sont disponibles sur mon canal.**

N.B. : Les exercices vous sont livrés pendant les séances des TD.

Éléments bibliographiques

Ayad, Abderrahmane (2024). *Lexicologie et sémantique. Étude contrastive des mots de l'islam*. Éditions Wamda, Jijel.

Ayad, Abderrahmane (2025). *Dictionnaire de linguistique et des sciences du langage.* Éditions science et pratique, Béjaia.

Bénoît, Habert, et coll. (1997). *Les linguistiques de corpus.* Armand Colin, Paris.

Eva, Soroli (2025). Cours de linguistique de corpus. URL: https://fr.scribd.com/document/949787726/Linguistique-de-Corpus-Cours-1

1. Définitions de base de la linguistique de corpus

1.1. Une première définition

Branche de la linguistique qui se propose d'extraire d'un corpus les connaissances linguistiques essentielles à l'enseignement des langues. Cette discipline a proposé un nouveau regard sur le langage en plaçant le sens dans le discours et non dans la pensée du locuteur.

1.2. Deuxième définition

La linguistique de corpus est une branche de la linguistique qui étudie le langage à travers des exemples contenus dans des textes réels. En particulier elle se propose d'extraire d'un corpus les connaissances linguistiques essentielles à l'enseignement des langues et à l'élaboration des dictionnaires.

2. Le corpus[1]

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

Le concept de corpus n'a pas reçu une définition unifiée ou stable[2] en sciences du langage. Il s'agit d'un ensemble d'énoncés que le linguiste délimite en vue d'étudier le système d'une langue ou des phénomènes langagiers. Ces énoncés constituant le corpus peuvent être soit oraux soit écrits.

Pour construire son corpus dans le but de mener une étude linguistique, lexicale, par exemple, le chercheur doit suivre quelques étapes nécessaires afin que son analyse soit cohérente et aboutisse à des résultats basés sur une démarche empirique.

Expliquant les critères de sélection d'un corpus d'étude, Georges Mounin précise :

« On choisit le corpus d'après des critères conceptuels : ensemble de textes théologiques, ou politiques, considérés comme tels à priori en dehors de tous critères linguistiques formels. Une fois constitué ce corpus, d'autre part, on exclut de l'analyse, par une deuxième opération purement conceptuelle, tous les termes qui sont jugés ne pas faire partie du champ étudié. »[3]

---

[1] Ayad, A. (2024). *Lexicologie et sémantique. Étude contrastive des mots de l'islam*. Éditions Wamda, Jijel, pp. 14-16.

[2] Garic N., et Longhi J., « L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique », Revue *Langage*, 2012/3 (n° 187), p. 3-11. [En ligne], consulté le 23 février 2021. URL : https://www.cairn.info/revue-langages-2012-3-page-3.htm

[3] Mounin G., *Clefs pour la sémantique,* Seghers, Paris, 1972, p. 135.

Puis dans la suite de cette explication, il vient à signifier un peu plus bas la nécessité de délimiter le corpus quand on a affaire à une analyse lexicale ; il note à ce titre :

« Ce qui signifie qu'un corpus ayant pour objectif une enquête lexicale spécifique ne pourra jamais être choisi au hasard, mais devra toujours être constitué à partir d'une décision conceptuelle, non linguistique : oralement, provoquer les sujets de la conversation, les thèmes traités, les centres d'intérêt ; par écrit, sélectionner des textes par sujets, thèmes ou centres d'intérêt définis conceptuellement » [4].

Ainsi, dans ce genre d'études fondées sur des corpus sélectifs, les conclusions lexicales et sémantiques auxquelles aboutira le travail d'analyse,– suivant un cheminement théorique et méthodologique exigé par la ou les disciplines-s dans laquelle/lesquelles on l'assoit, donc la lexicologie et la sémantique pour nous– sont le plus souvent d'une importance qui est telle que de nouvelles pistes de recherche se dessinent clairement à l'horizon ; c'est dire que les résultats obtenus, tellement la précision des données qu'ils apportent, la progression de l'examen est à même d'élargir le champ de la recherche pour appréhender de nouveaux aspects ou phénomènes linguistiques en relation avec l'étude déjà réalisée.

2.2. À quoi sert un corpus ?

Dans le domaine de la recherche, de la linguistique ou de l'enseignement, le corpus sert de base de données structurée pour analyser, décrire ou illustrer un phénomène. C'est un ensemble limité de documents (textes, enregistrements, images, vidéos) rassemblés selon des critères précis pour étudier un sujet donné.

En linguistique : Le corpus permet d'étudier des données réelles et authentiques (langue parlée ou écrite) pour analyser la fréquence de mots, les structures grammaticales, et les usages langagiers. Il sert à valider ou réfuter des hypothèses scientifiques. En ce sens, le corpus sert à explorer les différents secteurs de la langue (lexicologie, sémantique, onomastique, syntaxe, morphologie, phonétique, etc.).

Recherche (Sciences humaines et sociales) : Il sert d'échantillon représentatif pour étudier des faits de société, historiques, ou littéraires. Il permet d'extraire des informations, d'analyser des discours et de soutenir une argumentation.

[4] *Ibidem*, p. 136.

Enseignement (Pédagogie) : Le corpus de documents est un outil central pour mettre les élèves en activité, leur apprendre à analyser des sources et à construire des connaissances (notamment en histoire-géographie), et à traduire des séquences ou des énoncés, etc.

Traitement Automatique du Langage (TAL) : Les corpus numériques servent à entraîner des intelligences artificielles ou à créer des outils de traduction automatique.

2.3. Les ressources linguistiques numériques

Différents types :

- données de langage, de parole en format numérique
- outils et services pour le traitement de données linguistiques
- logiciels (préparation, collection, gestion, conversion, utilisation/reutilisation de données).

Exemples de ressources :

- Corpus écrits ou parlés, Lexiques, Grammaires,
- Ressources multimodales (vidéos, audios, images, gestes, codes sonores, etc.)
- Bases de données et dictionnaires terminologiques ou spécifiques à un domaine
- Ontologies, bases de données multimédia
- Systèmes de gestion et d'exploration de corpus
- Systèmes OCR (« Optical Character Recognition » ‘reconnaissance optique de caractères’)
- Pipelines / Systèmes de traitement de la parole (ASR: « Automatic Speech Recognition »),
- Systèmes de traduction automatique,
- Environnements d'annotation, de codage et d'évaluation de textes, etc.

2.4. Les types de corpus

En linguistique de corpus, le choix du type de corpus dépend de l'objectif de l'étude ‘analyse syntaxique, traduction, étude du lexique sémantique, lexicologie, lexicométrie, lexiculture, lexicomatique…’ interactions, etc.).

2.4.1. Le corpus monomodal (ou unimodal) : Un corpus monomodal est constitué de données qui n'utilisent qu'une seule modalité de communication

ou un seul type de média. Le plus souvent, il s'agit de texte écrit, mais cela peut aussi être de l'audio seul.

→ **Caractéristiques :** Le corpus unimodal présente généralement les caractéristiques suivantes :

■ **Homogénéité :** Les données sont de même nature (ex: uniquement des articles de presse, uniquement des transcriptions d'entretiens oraux).

■ **Objectif :** Il permet d'étudier en profondeur une structure langagière spécifique sans interférence d'autres modes (visuels, gestuels).

■ **Exemple :** Le corpus Frantext (corpus de textes écrits littéraires et scientifiques) ou une collection de posts d'un auteur sur un réseau.

2.4.2. Le corpus multimodal : Un corpus multimodal intègre plusieurs modes de communication (textuels, sonores, visuels) pour analyser la production du sens dans son ensemble. Il combine généralement la parole avec des gestes, des expressions faciales, des mouvements du corps, ou le texte avec des images.

→ **Caractéristiques** : Le corpus multimodal présente principalement les caractéristiques suivantes :

■ **Complexité :** Parfois il est nécessaire de disposer des outils de transcription et d'annotation sophistiqués pour analyser le son, la vidéo, l'image, le texte.

■ **Étude de l'interaction :** Idéal pour l'analyse des interactions orales, des discours publicitaires ou des documents numériques, ou encore pour l'étude des corpus religieux (émissions, vidéos, podcasts, etc.).

■ **Exemple :** Enregistrements vidéo de conversations en face à face où l'on analyse le discours (parole) en même temps que les gestes de la main et les mouvements de tête.

2.4.3. Le corpus parallèle : Un corpus parallèle (ou aligné) est composé de textes originaux dans une langue source (L1, l'arabe pour notre situation) et de leurs traductions dans une ou plusieurs langues cibles (L2, L3., le français pour nous).

→ **Caractéristiques :** Le corpus parallèle offre globalement les caractéristiques suivantes :

■ **Alignement :** Les textes ou énoncés sont segmentés et mis en correspondance (alignés) au niveau des phrases, des paragraphes, voire

des mots.

■ **Utilisation :** Principalement utilisé pour l'enseignement des langues, la traduction automatique, la linguistique contrastive et l'étude des équivalences lexicales.

■ **Exemple :** Un corpus contenant l'original du Coran (en arabe) et une de ses traductions françaises (Hamidullah 2000, par exemple), alignée page par page.

2.4.4. Le corpus phrastique (ou de phrases) : Un corpus phrastique est une collection de textes ou d'énoncés qui sont sélectionnés, organisés et analysés principalement au niveau de la phrase. Contrairement aux corpus de discours, l'unité de base est la phrase isolée ou un petit groupe de phrases.

→ **Caractéristiques :** Le corpus phrastique offre principalement les caractéristiques suivantes :

■ **Structure :** Souvent utilisé pour l'étude de la syntaxe, de la grammaire, de la rhétorique, de la morphosyntaxe...

■ **Contexte :** Peut manquer de contexte discursif large.

■ **Exemple :** Un ensemble de phrases extraites de supports écrits variés (livre, journal, posts, etc.) pour étudier la position des composants de la phrase et leurs fonctions, leurs sens… Les phrases peuvent être incorporées dans des paragraphes ou isolées (slogans, phrasèmes…).

2.4.5. Le corpus lexicographique : Un corpus lexicographique est spécialement conçu pour l'élaboration de dictionnaires ou l'étude du lexique. Il vise à être représentatif de l'usage d'une langue pour identifier les sens, les collocations (mots qui vont souvent ensemble) et les fréquences des mots.

→ **Caractéristiques :** Le corpus lexicographique présente généralement les caractéristiques suivantes :

■ **Taille :** Généralement très volumineux pour capturer les néologismes et les usages rares.

■ **Diversité :** Comprend des genres variés (écrits, oraux, techniques, littéraires).

■ **Exemple :** CNRTL (Centre National de Ressources Textuelles et Lexicales).

2.5.Un corpus hautement spécialisé : le corpus religieux[5]

En linguistique et sciences du langage, le chercheur se trouve face à une exubérance de choix de corpus de différentes natures. Parmi ces derniers, figurent les corpus religieux. Mais il en demeure, à notre avis, que l'un des plus riches en matière à explorer linguist-iquement, si ce n'est le meilleur, soit un corpus que l'on dépouillera du Qour'ên (« Coran »). Et, nous faisons ici allusion aux traductions françaises bien évidemment. Ainsi, de la centaine de traductions qui existe aujourd'hui, nous proposons tout particulièrement celle de Muhammad Hamidullah. Celle-ci porte le titre *Le Noble Coran et la traduction en langue française de ses sens*, sous son édition du Complexe Roi Fahd pour l'impression du Noble Coran, supervisée par le Ministère des Affaires Islamiques, des waqfs, de l'Appel et de l'Orientation, au Royaume d'Arabie Saoudite. Elle est publiée en 1420 H. (1999 G.). C'est une traduction du Texte Sacré de l'islam, le Coran, en arabe *El Qour'ên* qui est, à reprendre la définition du coronologue Nourredine 'Atr (*'Ouloûm El Qour'ên El Karîm* (Les sciences du Noble Coran), 1993 : 10) la Parole d'Allâh (Dieu) révélée au Prophète Mouhammed, qui est écrit dans des *Maçâhif* (exemplaires de ce Qour'ên), et transmis par des voies de narration multiples. Par sa récitation l'on accomplit un acte d'adoration, et il est inimitable ne serait-ce qu'une une seule sourate. Le Coran étant le dernier Message Céleste révélé sur terre, est de portée universaliste.

> Allâh l'a révélé pour toutes les générations humaines dans tous les pays jusqu'au Jour de la Résurrection. Allâh s'est Lui-même chargé de le préserver, car la fonction de ce Livre ne s'arrêtera qu'à la fin de la vie de l'humanité sur terre. » (S. El Fewzên, *El Irchêd Ilê Teshîh El I'tiqâd* (Le guide pour corriger la croyance) 2005 : 175). […].

Cela étant, les corpus religieux de l'islam peuvent être appréhendés sous l'angle d'une panoplie de disciplines linguistiques, telles que la lexicologie, la sémantique, l'analyse du discours en s'appuyant sur les exégèses de références (Tabarî, Ibn Kèthîr, etc.), pragmatique, énonciation, études dialogiques, interacti-on, terminologie, etc. Le chercheur optera alors pour une méthode de collecte de son corpus en prenant en compte la discipline dont il se servira des outils dans son étude.

> Il existe différentes modalités de constitution des corpus et une pluralité de méthodes et de démarches, qui posent un certain nombre de questions d'ordre épistémologique et nient la disponibilité d'un modèle idéal, qui serait d'emblée représentatif et herméneutique. Un

[5] Ayad, A. (2025). *Dictionnaire de linguistique et des sciences du langage.* Éditions science et pratique, Béjaia, pp. 104-117.

corp-us est nécessairement un construit, les données qu'il propose ont fait l'objet d'une sélection préalable. (Garic N., et Longhi J. : 2012).

Le corpus devra s'étaler de ce fait sur des énoncés-versets* variés du Noble Coran traduit en français, touchant à des thèmes et des items pluriels, tels que par exemple la croyance, l'adoration avec ses différentes formes (prière, jeûne, pèlerinage, etc.), la morale, les Noms divins, le comportement, la vie terrestre, l'au-delà, les bonnes et mauvaises œuvres… Une telle démarche permettra de ressortir les particularités des faits linguistiques à étudier. Il s'agit de ce point de vue de se fonder sur une linguistique de corpus qui serait plus à même de conduire à des résultats empiriques, en cheminant un itinéraire expérimental axé sur l'observation, la description et l'explication ; ces trois étapes étant incontournables dans toute analyse linguistique. Mettant au clair l'importance d'une telle démarche, François Rastier écrit :

> La linguistique de corpus pourvoit la linguistique d'un domaine où elle peut élaborer des instruments et définir une méthode expérimen-tale propre : elle ouvre aussi des champs d'application nouveaux et engage un mode spécifique d'articulation entre théorie et pratique. (...) On a trop souvent réduit les langues à des dictionnaires et des grammaires, voire à des syntaxes. Il faut cepen-dant tenir compte, outre du système, des corpus (corpus de travail et corpus de référence), de l'archive (de la langue historique), enfin des pratiques sociales où s'effectuent les activités linguistiques. Pour l'ess-entiel, une langue repose sur la dualité entre système (condition nécе-ssaire mais non suffisante pour produire et inter-préter des textes) et des corpus de textes écrits ou oraux. (F. Rastier, « La linguistique de corpus et refondation de la sémiotique des textes » 2016 : 20-22).

3. Comment construire un corpus ?

→Principes et critères

- Le type de corpus (définir la nature du corpus d'atude).
- La représentativité (la variété des textes, d'auteurs, de sources, de typologie textuelle, de régions géographiques.
- Faisabilité : temps et matériels à disposition.
- L'interchangeabilité d'un corpus : corpus réutilisable.
- Taille de corpus.
- Annotation et étiquetage.

Remarque : Dans toutes présentation des résultats de recherche, il faut impérativement bien décrire le processus de construction du corpus.