e-ScriptaRomanica
vol. 12 (2024): 11-27
ISSN 2392-0718
https://doi.org/10.18778/2392-0718.12.02
COPE Member since 2021 JM14948

Le corpus en sciences du langage, un lieu de vérification des enjeux langagiers

The corpus in language sciences, a place for verifying language issues

Salem Ferhat
École Normale Supérieure de Ouargla, Algérie
https://orcid.org/0000-0003-0848-3142
ferhat.salem@ens-ouargla.dz

Résumé : L'objet de l'article aborde la question du corpus en sciences du langage. Il met en exergue les principaux aspects relatifs à la constitution d'un corpus bien réfléchi pour servir d'assise théorique. Cet exposé est une synthèse d'approches permettant aux chercheurs débutants de constituer, de délimiter et d'explorer certains aspects du langage afin de lancer une recherche scientifique fondée. Il montre que les résultats d'une recherche en sciences du langage devraient découler des données tangibles d'un corpus pour que soit formulée toute nouvelle théorie, ou pour qu'une théorie soit mise en cause, voire infirmée. La matière en termes de contenus fait appel aux avis des spécialistes de corpus en ce qui concerne la définition du rôle du corpus dans les études linguistiques, de son choix, de sa délimitation, de son homogénéité, de sa représentativité, de la nature de ses composants et de ses contraintes. Cet article montre au fur et à mesure les principales facettes à prendre en compte dans la constitution du corpus afin de permettre aux chercheurs de garantir la fiabilité des résultats.

Mots-clés : corpus, sciences du langage, choix et limites, homogénéité et représentativité, fiabilité des résultats.

Abstract: The subject of the article addresses the question of corpus in language sciences. It highlights the main aspects relating to the constitution of a well-considered corpus to serve as a theoretical basis. This presentation is a synthesis of approaches allowing beginning researchers to constitute, delimit and explore certain aspects of language in order to launch well-founded scientific research. It shows that the results of research in language sciences should be drawn from the tangible data of a corpus before formulating any new theory, invalidating or calling into question another. The material in terms of content calls upon the opinions of corpus specialists with regard to the assigned definitions, the role of corpus for linguistic studies, its choice, its delimitation, its homogeneity, its representativeness, the nature of its components and its constraints. This article gradually shows the main facets to take into account in the constitution of the corpus in order to allow researchers to guarantee the reliability of the results.

Keywords: corpus, language sciences, choice and limits, homogeneity and representativeness, reliability.

Received: 03/03/2024; Reviewed: 15/05/2024; Accepted: 15/07/2024

> Un corpus doit-il répondre aux critères de pertinence pour justifier son usage en tant que lieu de vérification d'une hypothèse de travail ; il doit répondre aux critères de représentativité, d'homogénéité et de différence afin de rendre possible la mise en série des données et la généralisation/théorisation (Cislaru & Sitri, 2012, p. 61).

Introduction

Toute étude de caractère pratique part d'un corpus. Enchaîner des propos sur un point précis pour en identifier la nature, le fonctionnement, les spécificités, l'emploi, le rôle, la valeur et l'effet, constitue la matière et l'objet sur lesquels s'appuie tout travail scientifique en sciences du langage. En unité linguistique, micro ou macro, c'est en la matière qu'on pourrait dégager ce qui constitue une unité à part entière, déceler l'utilité d'un élément, valoriser sa présence et justifier scientifiquement sa cause. Le corpus est la matière première d'une « représentativité raisonnée » (Treffort, 2014) servant à justifier une cause. Sa constitution cible des données, en élimine d'autres, délimite le cadre de l'investigation. En fait, décrire et analyser un corpus, un point de langue, permet de définir la règle tout en conservant l'exception car c'est grâce à cette dernière que se confirment les définitions.

En effet, les propos déclarés sans fondement scientifique ne resteront que des déclarations conceptuelles, des hypothèses et des données non crédibles. Passer par et à travers des corpus, des énoncés achevés, authentiques, résultant de vraies situations de communication mènera le chercheur en langue à découvrir le fonctionnement des occurrences d'une unité linguistique et son effet en discours. C'est là où réside la dimension empirique des recherches qui permet en dernier ressort de déterminer une loi, d'en confirmer ou d'en infirmer une autre. Toutefois, et avant qu'un résultat s'impose comme étant une donnée de recherche, le corpus devrait se projeter et se soumettre aux contraintes de choix, d'homogénéité, de représentativité et de conditions de pertinence et signifiance de la matière afin que les résultats soient crédibles et fiables. C'est la raison pour laquelle nous avons voulu que cet article soit directif et propose un ensemble de données à interroger avant d'entreprendre un travail en sciences du langage. L'article est destiné en premier lieu aux étudiants en master sciences du langage pour leur servir d'aide dans l'élaboration de leurs données de départ. La matière de cet article est donc un exposé des principaux aspects relatifs à la constitution d'un corpus bien réfléchi pour servir d'assise empirique.

1. D'abord qu'est-ce qu'un corpus ?

Le corpus est une donnée brute, préalablement existante sous une quelconque forme. Parfois, il est le résultat d'une enquête, d'un questionnaire, d'une entrevue, d'une observation. Le corpus correspond à un ensemble, à un recueil et à une catégorisation de documents, regroupés suivant une optique précise et un principe de classement. Un corpus est un recueil de pièces ou de documents se rapportant à la même matière, à la même discipline ou à la même doctrine. Il pourrait se construire de mots, de locutions, de collocations, de phrases, d'énoncés, de textes de différentes tailles, d'ouvrages, de bases de données lexicales et textuelles, d'images, d'inscriptions tracées sur la monnaie, les rochers, les murs et les anciens papiers, de documents oraux (témoignages enregistrés ou transcrits), et de beaucoup d'autres choses encore.

Le corpus recouvre des documents relatifs à une discipline, réunis en vue de leur conservation. Ce principe de constitution pourrait concerner des ordres thématique, spatio-temporel, linguistique, caractéristique, catégorique, intervalle, formel, non formel, ou autres. Il correspond à un ensemble fini d'énoncés écrits ou enregistrés, constitué en vue de leur analyse linguistique. Pour Rastier,

> Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés de manière théorique réflexive en tenant compte des discours et des genres et de manière pratique en vue d'une gamme d'applications (Rastier, 2004, p. 2).

Selon le CNRTL[1], le corpus est un recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc. et en linguistique, selon la même source, il correspond à un ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique comme par exemple le corpus des textes parus d'un journal, d'une revue; un corpus littéraire; le corpus du vocabulaire français. Pour Dalbera,

> Dans les sciences du langage, un corpus est un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique. Il renvoie à une collection de textes présentant une certaine unité de genre ou bien d'époque (Dalbera, 2002, p. 1).

Le corpus ne se fonde pas sur la quantité des données, il vise plus la fiabilité en termes de sélection des données d'après des critères scientifiques afin que l'analyse donne des résultats fondés. Dans ses travaux sur le corpus, Tognini-Bonelli (2001) distingue deux approches du corpus. L'une considère le corpus comme une donnée basique, *corpus-based*, d'après laquelle une théorie, ou une hypothèse, se vérifie et dont l'objectif est de valider, de réfuter ou encore de remettre en cause cette théorie ou cette hypothèse. Cette approche fait du corpus un objet. L'autre approche considère le corpus comme étant une suite d'unités achevée pour identifier l'emploi et l'effet d'une unité, laquelle n'est donc pas envisagée dans son existence autonome en mot sans contexte. Cette approche se qualifie de *corpus-driven*, elle fait du corpus un moyen, elle est inductive car l'explication d'une question est issue d'un fait, elle tire le sens à partir des séquences macro.

2. Du traitement automatique du langage (TAL) aux bases de données

D'abord, « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (Habert, 2000, p. 1). Le corpus relève de données déjà prêtes ou de données à chercher, à sélectionner ou à reconstituer pour en faire une matière à étudier. Il pourrait concerner des écrits de toutes sortes ou encore se constituer d'un seul document dont les composants viennent d'une série d'autres documents. Le montage de cette série est tributaire du point que le chercheur voulait problématiser et le soumet à l'expertise scientifique.

Loin de toute description théorique et superficielle faisant appel à des exemples de corpus non authentiques, qui ne sont pas le résultat d'interactions authentiques, autrement dit des corpus construits et non achevés, les travaux linguistiques ont actuellement recours à des données informatisées. Ces travaux tirent du

[1] Centre National des Ressources Textuelles et Lexicales https://www.cnrtl.fr/definition/corpus (8/06/2021).

traitement automatique du langage des corpus fiables constituant la réalité de la communication humaine.

Le TAL trouve dans les données brutes une source pour constituer des données de corpus. Il sert de technique par laquelle le chercheur procède au choix, au tri et à la constitution de son corpus d'étude. En fait, de peur que le chercheur ne tombe dans le choix subjectif de données, les techniques de TAL ont permis de proposer des données neutres représentant la réalité des états d'une question à traiter. Selon Mellet,

> Le développement des bases de données informatisées, le formalisme obligé de leur structure, le questionnement sur l'échangeabilité des données, sur la standardisation des étiquettes et des formats associés ont sans doute ouvert la voie à une prise de conscience plus aiguë du phénomène (Mellet, 2002, p. 3).

De la sorte, et sans reconstitution de sa part, le chercheur, en linguistique par exemple, ne confrontera que les données authentiques et naturelles d'une réalité linguistique. Dès lors, le chercheur pourrait, selon la matière qu'il veut traiter, trouver, sur tel genre de texte ou telle unité linguistique ou autres, grâce à cette méthode scientifique de sélection et de collecte, des corpus complets et représentatifs pour la crédibilité de sa recherche.

Pour une étude linguistique, le chercheur pourrait regrouper, à travers les œuvres littéraires de tel siècle, de telle langue, les occurrences de tel adjectif en matière d'emploi afin d'étudier les effets de ce dernier sur les énoncés. Un autre chercheur pourrait aussi s'intéresser à l'étude de la préposition *à* dans un corpus de collocations relevant de telle langue. Un troisième chercheur pourrait penser à l'étude de la représentation schématique dans les articles scientifiques en didactique ou en linguistique. Un autre tiendrait des conversations des Français un corpus micro pour décrire la valeur d'emploi de *en fait,* massivement présent dans les échanges oraux. Dans ce même ordre d'idées, un autre encore prendrait l'adverbe *normalement* comme unité à étudier pour identifier ses valeurs d'emploi dans le contexte algérien car cet adverbe ne signifie pas seulement l'usage normal mais il acquiert encore le sens de *en principe*, d'*au lieu de,* de *ce qui est attendu*, de *ce qui est supposé,* de *ce qui est logique* et de *ce qui devrait être*. Dans ce même contexte, *normalement* fait partie d'un énoncé étant une cause logique du constat d'un désordre, d'une anomalie et non d'une chose bien arrangée. En fait, toute étude en linguistique part d'un point de langue, d'un aspect précis et d'un angle de vue.

La collecte des éléments de corpus pourrait se faire de plusieurs manières : par le recours aux données matérielles et tangibles à partir des traces écrites comme les œuvres d'un même auteur pour décrire son style et découvrir sa manière d'écrire, à partir des écrits sur tel évènement (les attentats du 1er septembre et les titres des unes de journaux), par l'enregistrement ou l'écoute des propos sur une question donnée et par beaucoup d'autres pratiques. Selon Moirand (2018), un chercheur en linguistique pourrait, grâce à ses lectures ou à ses déplacements, muni d'un carnet ou d'un autre support, recueillir et constituer des corpus à partir de ce qu'il entend dans les différents échanges langagiers en situations et en rapport avec ce qu'il veut exactement étudier.

3. Pourquoi s'appuyer sur un corpus pour étudier telle question ?

Faire appel à un corpus sert d'assise pour toute étude pratique. Le corpus constitue un véritable modèle pour la description des données sur lesquelles porte une recherche puisque « le corpus n'existe pas en soi, mais dépend (…) du positionnement théorique à partir duquel on l'envisage » (Charaudeau, 2009, p. 37). C'est pourquoi le corpus est indissociable de la théorie, et toute théorie est le résultat d'un corpus bien déterminé, voire délimité. Pour le même auteur,

> Les sciences du langage font donc partie des disciplines de *corpus* : rassemblement de données linguistiques (sous forme de textes écrits ou oraux, de documents divers, d'observations empiriques raisonnées ou d'enquêtes provoquées) que l'on constitue en objet d'analyse. Dès lors se pose la question de savoir quelle est la nature de ces données (Charaudeau, 2009, p. 39).

Pourquoi travailler sur des corpus ? Rassembler des faits de langue, sélectionner des unités linguistiques et les combiner avec d'autres pour vérifier leur compatibilité, repérer des unités linguistiques en fonction d'une situation de communication, identifier les coexistences de certaines unités sur le même syntagme, ainsi que d'autres modes organisationnels ; le chercheur en linguistique aura la possibilité d'apporter des descriptions à la grammaire d'une langue, de perfectionner les définitions dictionnairiques ; il pourra tester des hypothèses par l'empreinte réelle du langage humain, observer et analyser avec finesse les phénomènes langagiers, rendre crédible des résultats loin d'un corpus linguistique basé sur l'intuition ; le chercheur pourra en outre catégoriser grâce aux usages des données lexicales dans la chaîne parlée en raison de l'automation qui lui permet d'avoir une taille suffisante recouvrant toutes les occurrences possibles à partir de données textuelles. On pourra ainsi éviter de tenir des propos généralisés sur un phénomène alors qu'ils sont susceptibles d'en concerner un autre. C'est là où on s'écarte du caractère scientifique de la recherche de sorte que ces propos ne seraient pas fiable. Dans les études linguistiques, travailler sur des données de corpus servira, selon l'objectif du chercheur, à révéler les spécificités d'un point de langue par rapport aux occurrences et à la rencontre de ce même point comme paradigme dans les différentes combinaisons syntagmatiques. Le corpus permet également de valider une hypothèse avancée ou de l'infirmer avant toute décision de généralisation car « le corpus de textes constitue l'un des lieux les plus favorables à l'observation des réalisations de la langue pour la linguistique » (Comby ; Mosset, 2016, p. 7).

Travailler sur un corpus permet d'élaborer la théorie sur une donnée. À titre d'exemple, comme le constate Vetulani (2000, p. 323),

> Si un mot ne se comprend que dans le cadre d'une phrase, ceci signifie qu'une entrée de dictionnaire ne doit pas être constituée d'une unité lexicale, mais d'un niveau minimal d'analyse (d'une phrase élémentaire) permettant de rendre compte du fonctionnement d'une unité. Grosso modo il s'agit de ne pas séparer le lexique de la grammaire. Ce n'est malheureusement pas le cas de la plupart des dictionnaires.

Mais d'après cette remarque, si on la prend en considération, il faudrait des dictionnaires de grands volumes : chaque unité lexicale donnerait lieu à une infinité d'occurrences ce qui exigerait peut-être un volume à part car il s'agit bien des coexistences possibles d'une unité avec d'autres.

Dans le cas des études linguistiques, une première approche pourrait se focaliser sur l'aspect morphologique des mots s'ils sont mono, bi ou multi lexicaux dans le cas

où le chercheur linguiste s'intéresse à l'aspect formel des mots. Il pourrait par la suite inventorier tous les mots composés d'une langue, ou seulement une catégorie de mots, pour en connaître les règles fondatrices, à titre d'exemple ceux des noms d'agent relatifs aux métiers faisant le lien entre la spécialité et le spécialiste. Pour mieux cerner son corpus, le chercheur en langue tente par une analyse lexico-sémantique de décrire la valeur ajoutée à la base par l'introduction préfixale de *en* par une opération de listage pour voir si *en* dans *enquêter*, *entêter*, *endetter*, *enterrer*, *enrhumer*, *enlever*, *enrayer*, *enchaîner*, *enraciner*, *enrailler*, *enfiler*, *ensabler*, *enrouler*, *entourer*, *encaisser*, *enfariner, enfourcher, endommager, encoder, endosser, encercler, enrichir, engrainer,*et d'autres verbes encore produit, ou non, le même effet. En ce sens, le linguiste procède à la fois à une approche diachronique et synchronique pour mettre en évidence et décrire l'évolution de la langue car certaines langues suivent l'évolution du monde, ce qui fait de la création lexicale un mécanisme permanent. Pour Dalbera (2002, p. 8),

> Le corpus du linguiste est a priori l'ensemble des faits sur la base desquels celui-ci entend conduire son analyse. Ce corpus est, au premier chef, de l'ordre des ***données brutes*** : il consiste en un certain nombre d'unités linguistiques recueillies selon divers modes et rassemblées.

4. Constitution du corpus : conditions et enjeux

4.1. Choix et délimitation du corpus

En linguistique, le corpus correspond à l'« ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique » (Le Petit Robert, 1989). Il s'agit bien de choix, de limite et d'étude. L'exhaustivité n'est pas sollicitée car elle n'assure pas forcément des résultats fiables.

On pourrait encore parler d'un corpus fabriqué dont la collecte est le résultat d'un assemblage d'unités venant de plusieurs contextes sans caractéristiques communes. Procédure qui s'oppose à la linguistique de corpus qui analyse le sens en contexte et non pas, à priori, une donnée de la langue.

En linguistique, pour choisir et délimiter son corpus, on devrait s'interroger sur un certain nombre de points. On pourrait alors répondre aux questions suivantes : Je vais étudier quoi ? Quel point de langue ? Des unités linguistiques isolées ? Une unité linguistique en contextes ? De quelle langue ? D'une langue littéraire ou d'une langue non-littéraire ? De quel siècle ? D'une étude contrastive ou monolingue ? S'agit-il de l'écrit ou de l'oral ? En synchronie ou en diachronie ? De quel registre s'agit-il ? Du langage formel ou informel ? De quelle catégorie de personnes ? Hommes ou femmes ? Un corpus clos ou exhaustif ? Relevant de quel contexte social ? Est-ce le cas d'un natif ou d'un non natif ? Et d'autres questions encore. Ce sont des interrogations préalables pour cibler l'objet de la recherche et éviter le glissement auquel conduit involontairement la curiosité vers d'autres points similaires non abordés par le présent papier de recherche. Sur ce point de la collecte, Charaudeau prenait en considération certains aspects. Pour lui, c'est

> Le problème qui concerne le recueil des données, recueil qui dépend du choix de la matérialité langagière (paroles orales, paroles écrites), du choix du support qui véhicule ces paroles en relation avec une situation de communication (pour l'écrit : lettres, rapports, journaux, tracts, circulaires, affiches, etc. ; pour l'oral : radio, télévision, réunions diverses, meetings, conversations du quotidien, etc.). Ce sont autant d'aspects qui ont des incidences sur la manière de recueillir des données : exploration du terrain,

> procédés d'enregistrement libres ou contraints, au su ou à l'insu des acteurs de parole, etc. (Charaudeau, 2009, p. 37).

Une hirondelle, [seule], *ne fait pas le printemps*. Le nombre et la récurrence sous telle condition, et non pas le hasard, sont importants pour donner des explications scientifiques à un phénomène. La question du hasard et du concerté importe : une photo prise d'une personne sans que cette dernière en soit avertie, ou au contraire, une photo préparée à l'avance pour qu'elle apparaisse à son avantage, ne donneront pas les mêmes résultats pour confirmer scientifiquement une donnée informative. De même, si on veut mesurer l'impact de l'état psychologique et sa manifestation dans le langage, on devra étudier la personne, à titre d'exemple par des interviews, en intégrant plusieurs variantes (en santé, en état maladif, en état de choc, dans l'exercice de son métier, à la retraite, en position d'opulence ou en manque,…). Nous disons cela pour montrer le rôle que joue la définition du corpus dans la production des résultats fondus. En ce sens et dans le cadre d'une étude onomastique par exemple, on pourrait se demander comment nous avons formé les patronymes et les toponymes. Ce même sujet pourrait être traité dans une perspective comparatiste entre deux ou plusieurs cultures différentes.

Défini comme étant « la forme maximale du contexte » (Mayaffre, 2010, p. 13), le corpus incarne bien les différents usages d'une unité linguistique dans une série de textes de même nature, du même genre. Il permet de décrire et d'interpréter un phénomène langagier dans le cadre d'une époque de pensée, d'un domaine d'écriture, de la bibliographie d'un auteur, etc.

Enfin, on pourrait dire que le choix du corpus reste l'affaire du chercheur-enquêteur qui veut savoir l'état de la chose dans un cadre de recherche pertinemment ciblé. Ce cadre constitue pour lui une base matérielle intelligemment motivée pour servir de source informationnelle et de vérité scientifique.

4.2. Corpus et représentativité

Si le travail porte sur un corpus-échantillon, le chercheur devrait d'abord délimiter les faits à étudier avant de passer à l'analyse. Travailler sur un corpus restreint ou étendu est un point de départ qui aura un impact sur les éléments définitoires d'une confirmation. En effet, produire une définition en tenant compte du résultat d'une analyse d'unités linguistiques choisies assure au propos de la confirmation finale le fait d'être diffusé comme donnée scientifique à partager. Dans d'autres cas, notamment si le corpus est étendu, l'analyse ne pourrait atteindre le degré de fiabilité qui en ferait le résultat sûr d'une cause. Pour illustrer ces deux cas de corpus, macro et micro, nous mentionnerons le fait d'effectuer une recherche sur les techniques rédactionnelles en français comme un intitulé faisant appel à plusieurs genres d'écrit, en comparaison du fait de travailler sur un seul genre, par exemple l'écrit journalistique ou plus précisément le fait-divers. Les résultats obtenus ne seront pas les mêmes. En d'autres termes, la pertinence du choix méthodologique d'une recherche requiert de préciser le cadre de la quête : on comprend ainsi la nécessité de travailler sur un corpus très limité et représentatif pour pouvoir identifier le fonctionnement d'un point de langue dans toutes ses facettes dans le même contexte, contrairement au corpus exhaustif qui n'affine pas la recherche et ne donne généralement que des résultats semblables et applicables à d'autres cas.

La limite du corpus, d'après Dalbera (2002), relève de la responsabilité du chercheur. Qu'il soit étendu ou restreint, sa clôture dépend de l'objectif de la recherche. La représentativité du corpus dépend des aspects à traiter par la recherche ; elle est étroitement liée à la validité de l'analyse. La collection des

données de corpus devrait s'organiser « intellectuellement et matériellement » (Treffort, 2014). Les éléments qui forment sa matière devraient constituer préalablement le cadre définissant les angles d'attaque des hypothèses de recherche avancées. Selon Dalbera :

> L'extrapolation qu'il convient de faire pour étendre les résultats de l'analyse de l'échantillon à la langue impose que cet échantillon ait un caractère représentatif. La clôture du corpus ne peut plus être aléatoire ni seulement d'ordre quantitatif ; des contraintes qualitatives viennent s'ajouter, le corpus est alors de l'ordre des ***données pertinentes***. Par ailleurs la décision de garder le corpus ouvert a pour corollaire l'implication plus franche du linguiste dans le modelage de celui-ci ; le corpus est alors de l'ordre des ***données construites*** (Dalbera, 2002, p. 8).

Mais,

> À l'opposé des corpus homogènes et exhaustifs se trouvent les ***corpus échantillonnés*** ; là, le problème se déplace : l'enjeu n'est plus celui de l'exhaustivité, mais celui de la ***représentativité***. Il s'agit alors de constituer des échantillons représentatifs d'une réalité plus large (Mellet, 2002, p. 2).

Cela dévoile en effet les occurrences et rend crédibles les résultats obtenus à partir d'un corpus. À ce propos, il est à noter que, parfois, la représentativité du corpus est nécessaire du point de vue logique pour étudier un point incontournable et décrire son état et, dans d'autres cas, la représentativité n'est que personnelle recadrée suivant le choix du chercheur, mais selon un choix justifié car « il est […] relativement aisé de délimiter un échantillon représentatif de données, à condition bien sûr, d'assumer les exclusions », ajoute Dalbera (2002, p. 3).

Confirmer une réponse scientifique fondée et illustrée par des cas concrets nécessite un choix de corpus ; un choix qui part de la problématique de recherche et auquel porte cette dernière. D'où il convient de s'interroger sur la taille de corpus micro ou macro, devant telles situations de communication, dans tel genre de textes, corpus restreint ou étendu, relevant du même ou de plusieurs auteur(s), …En fait, du fait que plusieurs paramètres s'imposent avant de juger les résultats, « un corpus est donc toujours un ensemble jugé représentatif d'un tout, sans pour autant correspondre *stricto sensu* à ce tout » (Comby & Mosset, 2016, p. 10). En ce sens, le chercheur devrait prendre en compte les caractéristiques de son corpus : clos ou non clos, brut ou sélectionné, traité d'un brut. La question demeure toujours représentative car on suppose que « les régularités susceptibles d'être découvertes par l'analyste sont potentiellement récursives et donc qu'une analyse limitée à un sous-ensemble de faits peut être de nature à rendre compte de l'ensemble » (Dalbera, 2002, p. 2), mais est-ce bien toujours le cas ? Car dans certains domaines comme la psychologie où le chercheur se confronte à des cas individuels forts différents.

Éviter toute banalité et s'appuyer sur une donnée matérielle rendrait scientifiquement une telle recherche fondue et raisonnée. Le corpus délimité rend optimale la fiabilité des résultats et s'approche au maximum de la vérité d'un fait en comparaison d'explications purement théoriques et fantaisistes.

4.3. Nature de corpus

En général, les principaux critères de classification des corpus s'attachent à la nature des matériaux constitutifs, il peut s'agit en linguistique, à titre d'exemple, de textes complets ou limités à des unités de langue inférieures comme les phrases, les mots,

les phonèmes, les morphèmes, ou autres selon l'unité que le chercheur veut mettre en question.

On pourrait s'interroger sur l'objet concerné par l'étude s'il s'agit des unités macro ou micro ; des phonèmes, des morphèmes, des unités, des collocations, des énoncés, des séquences ou des textes. Aussi, des données de corpus s'il s'agit de données brutes, pertinentes ou construites. Autrement dit, si celles-ci sont authentiques ou élaborées. En ce qui concerne le statut d'une langue, on devrait préciser si la question de recherche porte sur une langue maternelle, seconde ou étrangère.

D'après Fray (2011), cité dans Magnani (2017), le corpus est « un rassemblement de textes ou une collection de textes regroupés sur la base d'hypothèses de travail en vue de les interroger ». Une définition qui amène à s'interroger sur la notion de texte. S'il s'agit du travail d'un historien cela pourrait se comprendre car ce dernier part des documents et des traces pour expliquer une question comme dans le cas de la numismatique. Dans le cas du linguiste, le *texte* ne correspond pas pertinemment à la définition du mot *corpus* s'il est devant un travail sur la ponctuation ou sur l'adjectif par exemple, on parlera de *texte* comme environnement linguistique qui met à la disposition du chercheur un maximum d'emplois et de reprises d'une unité comme l'adjectif dans une série de textes de même genre.

Le moment de la collecte constitue une première condition car les résultats des données de corpus reflèteraient l'état de la chose en fonction de la spontanéité ou de l'avertissement préalable des informateurs (informateurs avertis vs non avertis). À ce propos, nous notons l'impact qu'engendre la variété des conditions puisque les pratiques ordinaires et spontanées dans les échanges langagiers manifesteraient autrement le langage que lorsque les interlocuteurs se trouvent dans des situations de communications officielles (examen, écrit administratif, clauses de contrat ou convention, etc.). Cela constitue donc une condition expérimentale pour servir à la vérification des hypothèses avancées et pour enfin évaluer et juger les résultats.

À noter que des textes écrits, des témoignages oraux (enregistrés ou transcrits), selon des thématiques, des situations de communication, relevant d'un contexte social et produits à un moment donné, pourraient donner à une telle étude un caractère scientifique et une sorte de crédibilité, c'est pourquoi afin qu'une analyse devienne fondée et raisonnée, l'échantillon ne pourrait en aucun cas être choisi aléatoirement.

Comme synthèse définissant pertinemment un tel corpus, le chercheur en linguistique et en rapport avec la question à étudier pourrait trouver sa voie en s'interrogeant sur un certain nombre de points comme le montre la représentation suivante :

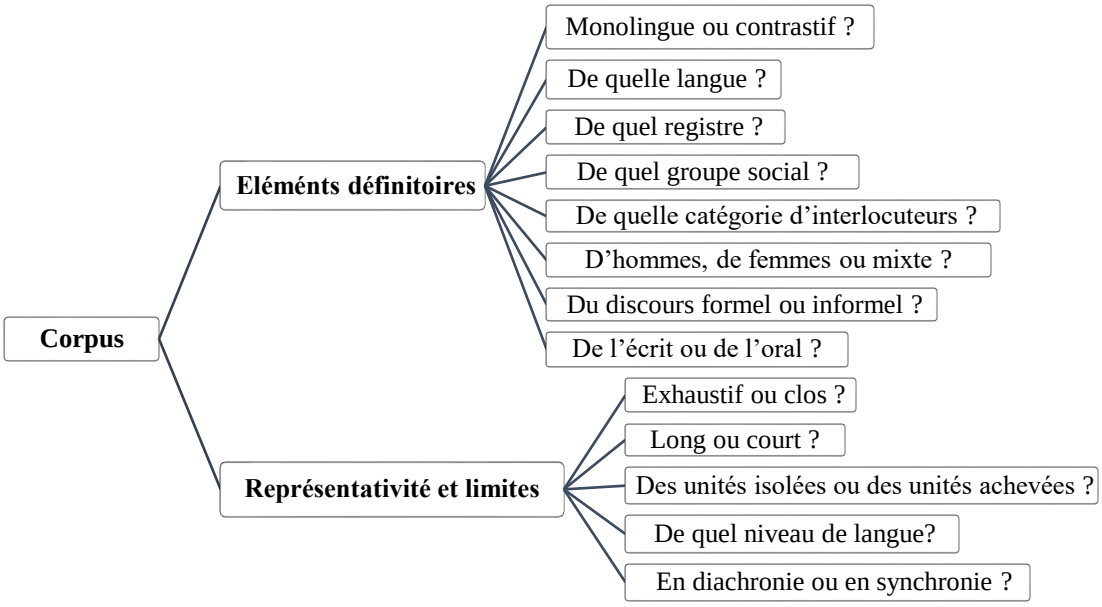

Schéma n° 01 : Définition et délimitation du corpus (étude linguistique)

Le chercheur en linguistique quand il procède à la description du corpus en termes de nature, devrait préciser dès le début l'objet de sa recherche. Il est appelé à dire à la communauté académique de quoi il s'agit. Il informe des éléments se rapportant à son corpus. Il met en exergue le fait que sa recherche n'est mise en œuvre que dans le cadre d'une langue, d'un corpus monolingue, ou le cas échéant d'un corpus contrastif ; il précise la langue en question, son registre et de quelle catégorie d'interlocuteurs il s'agit ainsi que leur appartenance ; il mentionne ensuite si ceux-ci sont d'ordre sélectif (hommes, femmes, écoliers, instruits, analphabètes,…). Il indique en plus si les éléments du corpus sont issus de telles circonstances, d'ordre oral ou écrit, d'un discours formel ou informel.

Pour ce qui est de la représentativité et des limites de son corpus, le chercheur informe des données de son corpus en précisant si ce sont des données générales, exhaustives, ou au contraire issues d'un corpus bien déterminé, un corpus clos. En termes de volume, le chercheur devrait justifier son choix de travailler sur un corpus court ou long, sachant que recourir, parfois, à un corpus court et représentatif suffit pour justifier l'état de la chose, un phénomène langagier par exemple. Quant au point de langue faisant objet de l'étude, le chercheur est invité à préciser le niveau de langue sur lequel porte l'étude, s'il s'agit du niveau phonétique, lexical, sémantique, syntaxique,… ; il précise en outre si son objet concerne une unité linguistique autonome ou combinée avec d'autres en contexte et si cela est envisagé en diachronie ou en synchronie.

4.4. Corpus et homogénéité

Pour assurer logiquement l'obtention des résultats fondés, il faut que le chercheur détermine la matière de son étude pour qu'elle soit pertinente sans se contenter de données brutes non catégorisées. Le corpus, comme le décrit Rastier, « est structuré d'une part en fonction d'une typologie des textes, qui se reflète dans leur codage, et d'autre part, dans chaque utilisation, par des sélections raisonnées de sous-corpus » (Rastier, 2005, p. 31). Selon d'autres chercheurs, l'uniformité des éléments de corpus constitue une condition pour réussir une étude et c'est pourquoi,

pour certains d'entre eux, « la linguistique recommande donc une attention particulière aux genres pour la constitution et l'analyse de corpus, en tant que facteur déterminant pour l'homogénéité du corpus » (Pincemin, 2012, p. 17).

S'il veut s'assurer de ses résultats, le chercheur devrait travailler sur des données de corpus dont l'homogénéité constitue une condition de fiabilité. Morsel (2016)[2] exclut qu'un corpus hétérogène soit forcément incohérent et non représentatif car la cohérence du corpus devrait tenir, dans le cas d'un ensemble de documents de plusieurs univers, à la logique dans laquelle est entamée une recherche et cela dépend toujours de l'objectif de cette dernière. À titre illustratif, ce pourrait être un échantillon composé des réactions de femmes et d'hommes lorsqu'ils entendent soudainement le cri d'un animal comme l'âne, sans avertissement préalable, pour savoir comment ils réagissent les uns par rapport aux autres. Ce corpus, donc, malgré sa nature hétérogène, est en adéquation avec l'objectif fixé au départ puisqu' « en se conjuguant avec le critère de représentativité, le critère d'homogénéité ne se réduit pas à une exigence d'uniformité », comme l'affirme Pincemin (2012). Dans le cas d'une étude en linguistique, le chercheur précise ce dont il va passer à l'examen ; s'il prend la conjonction *et* comme objet d'étude, il pourrait l'appliquer aux titres composant le sommaire d'une thèse afin de dresser son bilan comme il pourrait aussi étudier cette même conjonction dans les phrases complexes coordonnées. En littérature, un chercheur a la possibilité d'explorer la façon dont une même thématique est traitée par des auteurs de différents horizons ou encore la réception d'une pensée occidentaliste chez des critiques orientalistes ou l'inverse : l'Islam dans la presse francophone ; l'image de la femme dans la publicité en France, …

La question de l'homogénéité implique de s'assurer du fait que les données langagières à étudier relèvent du même ordre car

> Les corpus sont constitués de productions attachées à un dispositif situationnel spécifique, en général doté d'une forte institutionnalisation : textes publicitaires ou textes journalistiques, par exemple. L'homogénéité du corpus est établie par le dispositif de production (Garric, 2012, p. 75).

À titre d'exemple, ce peut être l'étude de la valeur modale de l'imparfait dans le récit à travers une série, repère, d'œuvres littéraires. Un autre exemple, qui pose problème dans son apprentissage surtout à des non natifs, est celui des temps du passé en langue française, un casse-tête qui pourrait constituer une problématique de recherche en matière de choix et d'adéquation d'un temps au détriment d'un autre. Si cela ne pose aucun souci pour les natifs, les non natifs en revanche s'y perdent et se trouvent indécis devant un passé à quatre facettes : *Je dis*, *J'avais dit*, *Je disais*, *J'ai dit*.

Quant aux études contrastives, l'homogénéité prend des corpus d'ordres différents comme matière première pour le lancement d'une recherche. Ces études établissent des rapprochements pour vérifier les corrélations et les fonctionnements d'un phénomène langagier au sein d'une famille de langues ou dans des genres de discours ressortissant d'une même langue. À ce titre, un chercheur pourrait envisager d'étudier le fonctionnement de la ponctuation ou de l'adjectif épithète dans une

[2] Lors de la journée d'études sur le thème : « Qu'est-ce qu'un corpus ? », organisée par l'équipe des CBMA (Chartae Burgundiae Medii Aevi) le 7 novembre 2016 à Paris. Passage inspiré des propos de la conclusion finale des travaux sur le corpus où Joseph Morsel s'est interrogé sur la cohérence et l'homogénéité du corpus. Dans Eliana Magnani. Qu'est-ce qu'un corpus ? Compte-rendu de la journée d'études. 2017, https://irht.hypotheses.org/3187. ffhalshs-01610087f

approche comparative du texte littéraire et du texte non littéraire. Il est possible aussi de voir comment les intitulés de thèses prennent forme en sciences du langage en comparaison avec les thèses en littérature. Sur cette question d'homogénéité du corpus, la constitution de la matière recouvre une autre dimension qu'on qualifierait parfois d'hétérogénéité, notamment si on étudie les troubles liés à l'articulation d'un apprenant ; cette étude nécessite la prise en compte des propos de ce même apprenant devant des personnes différentes car le fait de parler devant sa mère, devant ses enseignants ou devant un haut responsable a certainement un impact sur sa manière de parler. Il se pourrait qu'il parle à l'aise et en toute assurance devant sa mère et devant certains enseignants et que devant d'autres, enseignants et responsables, il éprouve de la gêne, il soit perturbé et ne parvienne pas à dire ce qu'il veut dire. C'est pourquoi ce genre d'étude devrait se faire en intégrant la variable personnes-différentes. Cela prouve bien que l'hétérogénéité est parfois au service de l'homogénéité, « l'hétérogénéité du corpus n'est pas en soi mauvaise. Elle peut être liée au terrain d'observation linguistique choisi », affirme ainsi Pincemin (2012).

5. Exemples de corpus

Avant de choisir son corpus, plusieurs interrogations s'imposent au chercheur. *En face de quelle question de langue le chercheur se trouve-t-il ?* est le premier point à déterminer avant de préciser à partir de quelles données du langage il faut étudier cette question de langue. C'est en rapport avec l'objet et avec l'objectif de l'étude que le chercheur détermine le corpus à travers lequel il procède à l'enquête. À titre d'exemple, le chercheur pourrait opter pour un corpus contrastif qui mettrait en jeu la variante homme-femme pour montrer et expliquer les raisons des différences apparaissant au niveau des prononciations des deux genres.

Sans que les exemples[3] qui suivent soient exhaustifs, ils serviront aux chercheurs en sciences du langage pour constituer leurs corpus à partir de quatre questions-guide : Quel est mon sujet de recherche ? Quelle est ma problématique ? Quel est mon objectif ? Et quel serait donc mon corpus ? Les réponses à ces questions élucident la matière sur laquelle s'effectue la recherche. Elles permettent de préciser les données textuelles autour desquelles s'examinent les hypothèses pour pouvoir répondre à la problématique.

Si, à titre d'exemple, un chercheur voulait travailler sur:

[3] Proposés par nous afin de cibler le corpus recommandé pour un chercheur-débutant. Il ne s'agit que d'une illustration et d'une liste non exhaustive.

Tableau 1 : Quel corpus en rapport avec le sujet, la problématique et l'objectif ?

	Un sujet comme :	Sa problématique serait :	Et que son objectif est :	Il pourrait alors envisager un corpus :
1	**Le langage de la bienvenue dans les conversations des réceptionnistes d'hôtels, une lecture dans la structure du langage et la nature des paradigmes**	Comment se structure le langage de la bienvenue et de quelles natures relèvent ses composants ?	Définir la structure de la séquence conversationnelle du langage de la bienvenue des réceptionnistes d'hôtels et identifier la nature de ses unités composantes	Constitué de séquences de conversations marquant le discours à l'arrivée
2	**Emploi(s) et fonction(s) de la locution adverbiale EN FAIT dans la conversation des français**	Quels effets produit EN FAIT par rapport à ses positionnements dans le discours des français ?	Identifier la position et la fonction de EN FAIT dans la conversation	Constitué de séquences orales où la locution EN FAIT est récurrente
3	**Les intitulés de travaux de recherche et l'usage de la ponctuation**	A quoi sert-elle la ponctuation dans le cadre d'un intitulé de recherche ?	Découvrir comment la ponctuation ajuste la signifiance au niveau des intitulés de recherche	Constitué d'intitulés de thèses de doctorat en langue française (thèses en linguistique, littérature et didactique)
4	**La communication non-verbale dans l'interaction entraineur-joueurs lors des matchs de football. Quels signes pour quelles significations ?**	Comment, à travers le non verbal, un tel signe de l'entraîneur comporte-t-il un sens aux yeux de ses joueurs? Comment les joueurs traduisent-ils les signes non-verbaux de l'entraîneur ?	Décrire la pertinence de choix des signes non-verbaux et leurs expressivités	Constitué de séquences de matchs de football relatives aux moments d'échanges entraîneurs-joueurs
5	**Les revues scientifiques en Algérie, quelle(s) lecture(s) derrière le choix de la dénomination?**	Que connote-t-il le nom de la revue ? et qu'imite-t-il pour les lecteurs ?	Identifier les raisons derrière le choix des noms des revues scientifiques algériennes	Constitué de l'ensemble des revues de la plateforme numérique des revues algériennes ASJP (Algerian Scientific Journal Platform)
6	**Titres des romans algériens et la 1ère de couverture, quelle correspondance et quelle signification ?**	Quel rapport entretiennent-ils le titre et la 1ère de couverture choisie au niveau des romans algériens ?	Expliquer la raison de choix de la 1ère de couverture	-Les romans algériens d'expression française -Les romans maghrébins postcoloniaux
7	**Structure(s) linguistique(s) et effet du langage, cas des exhortations des mendiants**	Comment l'effet se produit-il dans le langage des exhortations des mendiants ?	Identifier la structure de l'exhortation à travers le langage des mendiants	Constitué de propos relevant d'appels et de demandes des mendiants
8	**Pour une analyse pragmatique du verbe dans le proverbe**	Le verbe dans le proverbe est le résultat d'un choix qui dépasse l'aspect purement sémantique.	Découvrir la dimension pragmatique du verbe dans le proverbe	Constitué de proverbes français
9	**L'interjection et les manifestations corporelles dans la conversation en français**	Rapport interjection-corps et signifiance dans la conversation.	Décrire la combinaison interjection-corps et sa manière de signification	-Constitué de pièces de théâtres -Constitué de séquences de films
10	**L'émotion dans les discussions d'internautes à travers les réseaux sociaux**	Comment l'émoticon se combine-t-il avec langage ordinaire et donne-t- il sens ?	Décrire la combinaison interjection-corps modalité	Constitué d'émoticons faisant partie des propos d'internautes

6. Analyse et interprétation des données de corpus

Le sens premier relève de la description dictionnairique en l'état isolé d'un mot et la signifiance est l'affaire de l'usage qu'engendre ce mot par rapport à son environnement linguistique. Pour Charaudeau (In Adam & Viprey, 2009, p. 16),

> Le discours n'est pas le texte mais il est porté par des textes. Le discours est un parcours de signifiance qui se trouve inscrit dans un texte, et qui dépend de ses conditions de production et des locuteurs qui le produisent et l'interprètent.

Dès lors, plusieurs conditions devraient être envisagées pour arriver à la bonne interprétation puisqu'en analyse du discours, le texte n'est pas envisagé en tant que produit clos qui se suffit à lui-même en termes de signifiance ; il convoque en effet d'autres renseignements qui ne sont pas apparents dans la matérialité texte, c'est pourquoi,

> Le corpus n'est pas seulement construit, comme dans la plupart des domaines de la linguistique, en fonction d'un objectif de recherche ; il est par ailleurs contextualisé et mis en relation avec des « conditions de

> production », avec des pratiques sociales, plus largement avec des extérieurs qui le déterminent (Cislaru & Sitri, 2012, p. 61).

À ce propos, il est à noter aussi que c'est à cause du contexte de production d'un texte que, parfois, « la clôture du corpus devient impossible à tenir », comme l'ajoutent Cislaru et Sitri (2012, p. 61).

L'étude de la langue se fait dans sa réalisation en langage et les linguistes partent toujours de l'unité *texte*, en considérant « le texte comme l'unité minimale et le corpus comme l'ensemble dans lequel cette unité fait sens » (Adam & Viprey, 2009, p. 16). Pour les linguistes, l'échange langagier entre les interlocuteurs se fait en grande partie à travers des suites de mots, en texte et pas en état de mots isolés. C'est pourquoi « une démarche de sciences du langage se doit d'aborder d'abord les *textes,* en tant qu'ils constituent les réalisations empiriques premières de l'ordre langagier » (Bronckart, 2008, p. 39). D'où le fait que l'étude même d'une unité lexicale devrait se faire en rapport avec son usage dans tel ou tel texte. Pour délimiter un corpus, il faut étendre son cadre et réfléchir à la contextualisation[4]. Cette dernière est définie « comme un processus par lequel le chercheur tente d'établir la pertinence d'une mise en relation entre un texte et un autre (ou plusieurs autres) et de leur regroupement au sein d'un corpus » (Capt & *al*., 2009, p. 129).

Plusieurs paramètres doivent être pris en considération dans l'analyse et l'interprétation du corpus. Le cadre spatiotemporel, les circonstances liées aux échanges langagiers, les interlocuteurs, la thématique de leur discours et d'autres encore sont tous des repères de signification. Le corpus devrait être considéré dans une large dimension, linguistique et contextuelle. Pour Charaudeau, c'est là

> Le problème qui concerne, à l'intérieur du matériau langagier, les catégories qui vont faire l'objet de l'analyse : grammaticales (connecteurs, pronoms, verbes, etc.), lexicales (par champs ou de façon aléatoire), syntaxiques (selon divers types de construction) ; mais aussi les variables externes à la production des actes langagiers, telles que les types de locuteurs, les dispositifs de communication, de même que les variables concernant le temps (l'historicité) et l'espace (les cultures) (Charaudeau, 2009, p. 38).

C'est pourquoi, selon Mayaffre, le cadre étendu de l'analyse devrait mener à savoir quelle signification nous donnons à tel mot. Pour le même auteur,

> On ne pouvait comprendre un mot sans la phrase et la phrase sans le discours, on ne pouvait comprendre le discours sans l'interdiscours, le texte sans le co-texte (sans même parler ici du hors-texte), c'est-à-dire aussi et de manière plus générale, le corpus sans le hors-corpus (Mayaffre, 2002, p. 5).

Ces considérations, donc, dépendent de l'intérêt et de l'objectif du chercheur, ce dernier fixe son corpus et passe à l'exercice, il assume son choix méthodologique et son angle d'attaque. En réalité, tout corpus n'est qu'une facette d'une possibilité de la langue et c'est la raison pour laquelle « assumer la singularité et la subjectivité d'un corpus ne vise pas à remettre en cause sa pertinence, mais bien à le confronter à la littérature existante et à d'autres sources ou corpus pour parvenir à consolider les résultats » (Comby ; Mosset, 2016). Il s'agit d'un prolongement qui fait appel aux circonstances et conditions de la production

[4] D'après la conception de Capt, Jacquin et Micheli (2009). Pour eux, la définition d'un corpus est soumise à l'interrogation des sphères de contextualisation *générique*, *auctoriale* et *thématique*.

du discours ; pour les linguistes tout acte est à resituer dans ses circonstances d'effectuation afin d'obtenir plus de renseignements sur sa raison d'être, sa cause à effet. Dans cette optique, Charaudeau montre l'impact de ce prolongement en contexte étendu en expliquant que :

> Dans cet élargissement progressif de la notion de contexte, apparaît une prise de conscience progressive, non seulement du rapport entre texte et tout l'environnement textuel qui peut s'y rapporter, mais aussi entre le texte et un « hors-texte » (parfois appelé *cotexte*), c'est-à-dire des données présentes dans les *conditions de production* de l'acte de langage (Charaudeau, 2009, p. 46).

Enfin, il convient de rappeler que l'interprétation d'un fait de langue ne peut se limiter à de simples données de corpus puisque la crédibilité scientifique impose au chercheur d'optimiser les conditions de la quête afin qu'il puisse formuler des lois.

7. De la description et l'analyse de corpus à la théorisation

Mesurer l'ensemble passe par mesurer une partie représentative. C'est grâce à la condition de signifiance qu'un corpus se constitue. En sciences du langage, les données du langage, des unités isolées ou des unités en usage, sont un préalable fondamental de l'analyse scientifique qui aboutit à toute théorisation sur la langue. Soumettre les données d'une langue à la description et à l'analyse constitue la base de toute théorisation même si parfois le chercheur pourrait faire l'inverse quand il veut savoir l'opérationnalité d'une théorie préexistante sur un autre corpus d'une langue différente, qui n'est pas la langue originaire dont est issue cette théorie.

Avant de passer à la théorisation à partir d'un corpus, observer, rechercher et décrire sont des phases indispensables dans la formulation des lois et la constitution des données scientifiques à vulgariser. Plusieurs interrogations se posent sur la représentativité du corpus pour pouvoir théoriser scientifiquement tel état d'une langue. Aussi, la validité de telle théorie et sa pertinence devraient délimiter le cadre référentiel de l'application et ses exigences pour qu'il y ait toujours le même résultat. Cette finalité est tributaire de certaines conditions étroitement liées au choix du corpus. Écarter certaines anomalies, celles qui font défaut par rapport à l'ensemble de données du corpus, sert en effet à bien cerner la règle à partir de laquelle s'instaure une théorie, « c'est finalement le corpus qui fait la théorie » (Dalbera, 2002, p. 9). S'arrêter donc sur tous les détails du corpus, le décrire et l'analyser avec finesse permet de trouver les analogies et les spécificités grâce auxquelles se confirment les règles.

L'objectif de travailler sur un corpus est, pour la linguistique, de trouver une base théorique à travers laquelle s'opère une validation empirique. Ce sur quoi grâce à cette dernière se fonde le jugement scientifique et la nouvelle théorie s'impose avec ses résultats sur les mêmes corpus. Théoriser consiste à assurer les mêmes résultats à partir des données de corpus plus ou moins stables sur d'autres éléments semblables et ce par la même démarche scientifique. C'est en fait un travail qui part d'un corpus pour en décrire le mode de fonctionnement, lequel sert de règle qui s'applique aux autres corpus identiques ou semblables.

Conclusion

La linguistique de corpus qui part des données du discours, mises en jeu dans les différentes manifestations langagières, constitue une réalité de la langue en fonctionnement et c'est sur elle que la description et l'analyse des faits de langage devraient s'appuyer si le chercheur voulait dresser un tel bilan.

Selon son choix, sa délimitation et sa représentativité, le corpus reste la seule matière qui pourrait révéler les enjeux langagiers et permettrait à l'analyste de donner l'explication à sa raison de combinaisons en discours. À cette fin, et afin qu'il représente l'ensemble, le corpus est soumis à des exigences formelles et représentatives. Ces dernières devraient porter sur son cadre d'étude en termes de délimitation des unités ou de séquences plus au moins longues à vérifier ; le chercheur veille aussi à la représentativité des données choisies par rapport à l'ensemble pour pouvoir tirer des résultats scientifiquement fondus et susceptibles de garantir la constance sur d'autres corpus de même nature.

Avant d'entamer l'analyse proprement dite sur un tel corpus, des questionnements portant sur la langue à étudier, l'unité concernée, le contexte d'usage, sa réalisation orale ou écrite, son caractère formel ou informel, et beaucoup d'autres paramètres encore, permettent de cibler pertinemment le corpus et servent à décrire logiquement sa nature. Le corpus, comme nous l'avons vu, ne se rattache pas à la quantité des données, il vise plus la fiabilité en termes de sélection des données d'après des critères scientifiques afin que l'analyse soit optimale et permette le fondement des résultats. Pour arriver à cette sélection des faits de langage en matière de délimitation, les bases de données et les logiciels de constitution de corpus, à travers le TAL (Traitement Automatique du Langage), offrent actuellement aux chercheurs la possibilité de travailler sur des corpus bien ciblés.

Qu'étudier exactement, à partir de quelles données du langage et sur combien de documents faut-il enquêter ? À ces interrogations s'en ajoutent d'autres portant sur la condition de signifiance, la représentativité et l'homogénéité, pour qu'un corpus se constitue et devienne une matière première prête à l'expertise car mesurer l'ensemble passe par mesurer une partie représentative. Bref, si toutes ces conditions sont réunies, les résultats deviennent fiables et serviront à la théorisation.

Bibliographie

ADAM, J.-M. & VIPREY J.-M. (2009). Corpus de textes, textes en corpus. Problématique et présentation. *Corpus*, 8, pp. 5-25.

BRONCKART, J.-P. (2008). Genres de textes, types de discours et degrés de langue. Hommage à François Rastier. Texto [en ligne]. Dialogues et débats. URL : https://www.revue-texto.net/index.php?id=86

CAPT, V., JACQUIN, J. & MICHELI, R. (2009). Les sphères de contextualisation. Réflexion méthodologique sur les passages de texte à texte(s) et la constitution des corpus. *Corpus*, 8, pp. 129-147.

CHARAUDEAU, P. (2009). Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus*, 8, pp. 37-66.

CISLARU, G., SITRI, F. (2012). De l'émergence à l'impact social des discours : hétérogénéités d'un corpus. Langages, Armand Colin, pp. 59-72.

CNRTL (Centre National des Ressources Textuelles et Lexicales). https://www.cnrtl.fr/definition/corpus [7/05/2021].

COMBY, É. & MOSSET, Y. (2016). *Le corpus à l'interface des humanités et des sciences sociales*. Dans *Corpus de textes : composer, mesurer, interpréter*. Lyon : ENS Éditions. http://books.openedition.org/enseditions/7341 ; DOI : https://doi.org/10.4000/books.enseditions.7341

DALBERA, J-P. (2002). Le corpus entre données, analyse et théorie. *Corpus*, 1. http://journals.openedition.org/corpus/10 ;
DOI : https://doi.org/10.4000/corpus.10

GARRIC, N. (2012). Construire et maîtriser l'hétérogénéité par la variation des données, des corpus et des méthodes. *Langages* 3,187, pp. 73-92.
DOI : https://doi.org/10.3917/lang.187.0073

HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In BILGER, M., editor, *Linguistique sur corpus. Études et réflexions*, number 31 in *Cahiers de l'université de Perpignan*, pp. 11-58. Perpignan : Presses Universitaires de Perpignan.

LE PETIT ROBERT. DICTIONNAIRE ALPHABETIQUE ET ANALOGIQUE DE LA LANGUE FRANÇAISE (1989). dirigé par Alain Rey et Josette Rey-Debove. Paris : Société du Nouveau Littré/Le Robert.

MAGNANI, E. (2017). Qu'est-ce qu'un corpus ? Compte-rendu de la journée d'études. https://irht.hypotheses.org/3187 [18/04/2021].

MAYAFFRE, D. (2002). Les corpus réflexifs : entre architextualité et hypertextualité. *Corpus*, 1. http://corpus.revues.org/11 ; DOI : https://doi.org/10.4000/corpus.11

MAYAFFRE, D. (2010). *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, mémoire pour l'Habilitation à diriger des recherches, Université Nice-Sophia Antipolis. http://tel.archives-ouvertes.fr/tel-00655380 [28/09/2017].

MELLET, S.(2002). Corpus et recherches linguistiques. *Corpus,* 1.
http://journals.openedition.org/corpus/7 ; DOI : https://doi.org/10.4000/corpus.7

MOIRAND, S. (2018). L'apport de petits corpus à la compréhension des faits d'actualité. *Corpus,* 18. http://journals.openedition.org/corpus/3519 ;
DOI : https://doi.org/10.4000/corpus.3519

PINCEMIN, B. (2012). Hétérogénéité des corpus et textométrie. *Langages*, 187, pp. 13-26.
DOI : https://doi.org/10.3917/lang.187.0013

RASTIER, F. (2004). Enjeux épistémologiques de la linguistique de corpus. *Texto !* http://www.revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Enjeux.html [10/09/2021].

RASTIER, F. (2005). Enjeux épistémologiques de la linguistique de corpus. In WILLIAMS, G. (éd.). *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, pp. 31-45.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. *Studies in corpus linguistics, 6.* https://irht.hypotheses.org/3187#identifier_3_3187 [11/02/2021].

TREFFORT, C. (2014). Le corpus du chercheur, une quête de l'impossible ? Quelques considérations introductives. *Le corpus. Son contour, ses limites et sa cohérence, Annales de Janua, Actes des Journées d'études.* http:// annalesdejanua.edel.univ-poitiers.fr/index.php?id=725 [28/07/2021].

VETULANI, G. (2000).Quelques exemples d'analyse des corpus en vue de la traduction. *Studia Romanica Posnaniensia*, 25/26, pp. 317-325.