LES CORPUS ET LEUR EXPLOITATION

Christophe Benzitoun & Paul Cappeau
(03-2025)

Pour citer cet article :
Benzitoun (C.) & Cappeau (P.), 2025, « Les corpus et leur exploitation »,
in *Encyclopédie Grammaticale du Français*, en ligne : http://encyclogram.fr
DOI ; https://nakala.fr/10.34847/nkl.********

0. INTRODUCTION

La constitution et l'utilisation de corpus n'ont cessé de croitre depuis les années 1960 dans le domaine de la linguistique française, avec une accélération au début des années 2000. En conséquence, la linguistique de corpus est devenue une approche importante et un grand nombre de recherches sont désormais menées et publiées dans ce courant. Elle ne se réduit pas à l'usage de corpus comme source d'une description. De nombreux paramètres sont discutés pour en déterminer les bonnes pratiques et les limites. L'informatique a joué un rôle majeur dans son essor.

Conçue comme une boite à outils faisant le point sur les pratiques et ressources existantes en linguistique de corpus, cette notice a une orientation volontairement pratique. De ce fait, nous développerons peu certains aspects théoriques que divers ouvrages – pour la plupart en anglais – abordent de façon approfondie (McEnery and Wilson, 2011). Si de nombreux livres d'introduction à la linguistique de corpus sont en anglais, il en existe tout de même quatre qui sont rédigés en français : Poudat et Landragin (2017) ainsi que Zufferey (2020), assez récents, et Habert, Nazarenko et Salem (1997) et Habert (2006), plus anciens.

Nous nous centrons sur la constitution et l'utilisation des corpus principalement avec une visée grammaticale. Pour une présentation des corpus dans le champ de la linguistique de l'interaction, se reporter à la [> Notice de Mondada & Pekarek-Doehler].

Dans la première partie, nous détaillons les origines et les problèmes définitoires de la notion de corpus. Dans la deuxième partie, nous présentons les balbutiements de la linguistique de corpus et discutons de la situation dans les années 2020. La troisième partie est dans la continuité des deux premières, avec une visée historique, et aborde les critiques qui ont été faites, notamment par le courant génératif, concernant le recours au corpus comme source de connaissance en linguistique. Les deux parties suivantes représentent le cœur de la notice et dressent, de manière détaillée, les paramètres à prendre en compte quand on travaille à partir de corpus. La dernière partie présente un bilan provisoire des risques qui pèsent sur la linguistique de corpus accompagné de quelques perspectives.

1. LA NOTION DE CORPUS

1.1. Histoire du terme

Le terme latin *corpus* a entre autres sens celui d'*ensemble structuré d'éléments*. Il est employé depuis l'antiquité dans la langue savante pour désigner des collections de textes (juridiques, épigraphiques, philologiques…). Le dictionnaire étymologique et historique du galloroman (ou *FEW*) de Walther von Wartburg permet de suivre l'évolution du mot *corpus* avant son utilisation par les linguistes (http://www.atilf.fr/FEW/). Le mot est utilisé en ancien et moyen français dans un contexte religieux (*corpus Domini* et *corpus Christi*), puis il prend le sens de « collection, p.ex. d'inscriptions, surtout de l'antiquité ». À la fin du XIX^{ème} siècle, on l'utilise en droit notamment à travers deux locutions : *corpus iuris* i.e. « collection du droit romain » et *corpus delicti* soit « fait principal, objet qui constitue la preuve du délit ». C'est la notion de *collection* qui est devenue centrale dans l'usage contemporain du mot *corpus*. Ainsi, en sciences humaines, « il s'agit d'un recueil, formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline. » (Mellet, 2002). Le terme est régulièrement utilisé aussi pour désigner l'ensemble des œuvres d'un musicien, d'un écrivain, d'un philosophe, d'un courant de pensée, etc.

L'apparition du terme chez les linguistes date de 1961 selon A. Rey (1992), avec le sens d'« ensemble d'énoncés servant de base à l'analyse » ; du milieu des années soixante selon Teubert (2009) ; ou des années 1970 si l'on suit Blanche-Benveniste (1997). On trouve également une courte section intitulée *Le « corpus »* (avec des guillemets) dans les *Éléments de linguistique générale* d'A. Martinet parus en 1960 (pages 30-31). On peut donc raisonnablement considérer que le terme est apparu au tout début des années 1960. Mais, en fait, selon Laks (2010 : 12), dès l'Antiquité, deux rapports distincts aux observables langagiers peuvent être dégagés, qu'il oppose en distinguant ce qu'il nomme les sciences de l'*exemplum vs* du *datum* : « Les exemples, construits ou repris des meilleurs auteurs, sont saisis comme les éléments d'une pédagogie de la rhétorique. Au contraire, les usages collectés sont vus comme la base de construction de grammaires descriptives. » L'approche contemporaine des corpus s'inscrit, si l'on suit cette analyse, dans une tradition bien plus ancienne. Auparavant, on utilisait déjà des corpus sans les désigner de la sorte.
De même, l'exploitation sous forme de concordances (à savoir l'ensemble des occurrences du mot ou du groupe de mots recherché en contexte) remonte, selon McCarthy et O'Keeffe (2010 : 3), au XIII^{ème} siècle :

> « La concordance est née d'un besoin pratique de spécifier aux autres biblistes, par ordre alphabétique, les mots contenus dans la Bible, ainsi que les citations de l'endroit et des passages où ils apparaissent. » (trad. Cappeau)

1.2. Les premiers corpus informatisés en linguistique

Au début des années 1960 paraissent les deux premiers corpus informatisés de langue anglaise, à savoir le *Brown Corpus* et le *Survey of English Usage*. Les deux sont composés de centaines d'échantillons de textes pour un total d'un million de mots

chacun, mais seul le SEU comporte également des transcriptions d'enregistrements de conversations spontanées. Un autre corpus informatisé, moins connu, est apparu en 1949 et conçu à des fins de traduction automatique (Léon, 2005) : il s'agit du *Rand Corpus* composé, entre autres, de centaines d'articles russes de physique et de mathématiques.
Du côté français, une entreprise similaire voit le jour à la fin des années 1950 et aboutira un peu plus tard au corpus informatisé du *Trésor de la Langue Française*, plus volumineux que les précédents, composé de 70 millions de mots et uniquement de textes intégraux. Ce corpus n'a cessé d'être enrichi au fil des décennies. Il se nomme désormais *Frantext*, est largement utilisé et compte désormais plusieurs centaines de millions de mots.
Il n'est pas évident de dire quel a été le premier corpus informatisé disponible pour des recherches en linguistique. Si l'on part du lancement du projet ou de sa finalisation, on ne parvient pas à la même conclusion[1].

1.3. Problèmes de définition et de terminologie

À l'heure actuelle, un corpus est généralement considéré comme étant une collection de textes écrits et/ou d'interactions orales authentiques regroupés pour un objectif précis. La majorité des corpus sont désormais informatisés et sont exploités à partir de la lecture des données ou bien, plus fréquemment, à l'aide de logiciels plus ou moins polyvalents. L'informatisation représente une évolution incontournable dans le domaine, mais cela ne remet évidemment pas en question le statut de corpus des données collectées avant l'ère de l'informatique.
Au-delà de la définition générale ci-dessus, le terme *corpus* possède trois sens spécialisés dans les travaux récents en linguistique française :

• Soit il s'agit d'un ensemble de textes qui va permettre au linguiste de rechercher des attestations qu'il va analyser par la suite. Dans ce sens, Rastier (2004) utilise la désignation « corpus de référence ». Dans la citation ci-dessous, le terme *corpus* est employé pour décrire une telle ressource :

> « Cette étude a été menée à partir du **corpus** nommé CORPAIX, recueilli entre 1977 et 1999 par ce qui constituait alors le Groupe Aixois de Recherche en Syntaxe (GARS), dirigé par Claire Blanche-Benveniste. » (Chanet, 2001 : 56)

• Soit il s'agit de données sélectionnées pour mener à bien un projet d'études, ce que Rastier (2004) nomme un « sous-corpus de travail en cours ». Il « peut ne contenir que des passages pertinents du texte ou des textes étudiés ».

> « Nous avons sélectionné dans l'ensemble des productions la partie la plus homogène et la plus représentative, qui est aussi celle où l'institutrice intervient le moins, les narrations (6000 mots environ). C'est cette partie qui constituera le **corpus** transcrit en annexe. » (Martinot, 2005 : 35-36)

[1] Pour avoir plus sur les premiers corpus informatisés, on peut se reporter à https://htl.cnrs.fr/le-saviez-vous-17/ (Léon, 2024) et aux références bibliographiques qui y sont mentionnées.

La différence entre ces deux premiers sens tend cependant à s'atténuer et il peut devenir difficile, voire peu utile, de les distinguer. Le chercheur analyse son corpus de travail étroitement lié au corpus de référence (réduit aux séquences exploitées) :

> « Ensuite, les catégories d'opérations proposées ci-dessous ne sont nullement exclusives les unes des autres. Elles ont même tendance à se superposer, et les différencier n'a pour but que de cerner un peu mieux la fonction que peut revêtir *quoi* en discours. Bon nombre d'exemples du **corpus** cumulent en effet plusieurs des phénomènes passés en revue ici. » (Chanet, 2001 : 69).

> « Le **corpus** que nous analyserons dans ce livre révèle une grande uniformité des phrases attestées chez tous les enfants. » (Martinot, 2005 : 2)

• Un troisième sens que l'on retrouve parfois est celui d'un ensemble d'énoncés servant de base à l'analyse. Dans ce cas, le corpus se réduit à une liste d'exemples. C'est ce sens que l'on rencontre souvent dans des travaux de morphologie :

> « En nous appuyant sur l'examen d'un **corpus** de plus de 6 500 dérivés d'anthroponymes » (Huguin *et al.*, 2023)

> « Leur recherche se fonde sur un **corpus** construit *ad hoc*, d'environ 300 mots-valises tirés de diverses sources relevant des différentes variétés diatopiques du portugais (européen, brésilien, angolais et mozambicain). » (Grossmann *et al.*, 2022)

Rastier (2004) défend une position très restrictive (peut-être trop ?) qui ne correspond pas toujours aux usages contemporains :

> « Notre définition suppose qu'un corpus n'est pas un corpus de mots (cf. à l'inverse le projet européen *Paroles*) ; ni un corpus d'attestations ou d'exemples (comme Frantext, dès lors qu'on n'a pas accès aux textes-sources) ; ni un corpus de fragments (comme le *British National Corpus*, qui ne contient aucun texte complet, mais un échantillonnage). »

L'objet d'étude peut exercer une influence déterminante sur la configuration du corpus, comme le précise Maingueneau (2005 : 73) :

> « Des unités comme « le discours raciste », « le discours postcolonial », « le discours patronal », par exemple, ne peuvent pas être délimitées par des frontières autres que celles qu'a posées le chercheur ; elles doivent en outre être spécifiées historiquement. Les corpus auxquels elles correspondent peuvent contenir des énoncés relevant de types et de genres de discours les plus variés ; ils peuvent même, selon la volonté du chercheur, mêler corpus d'archives et corpus construits pour la recherche (sous forme de tests, d'entretiens, de questionnaires…). »

Sur la diversité des définitions du terme *corpus*, on peut renvoyer aux ouvrages de synthèse qui seront cités dans cette notice (chaque ouvrage propose sa propre définition, souvent très proche de celle des ouvrages concurrents). C'est notamment ce qui explique pourquoi l'accent est souvent mis sur la présentation du corpus, sur la mise en évidence de son originalité et de sa cohérence et sur la disponibilité et l'utilisation de toutes les informations annexes (ou données secondaires ou *métadonnées*) qui permettent de situer et de contextualiser les ressources. La description documentaire de la ressource est une dimension importante de la linguistique de corpus.

Par exemple, il parait difficile de justifier un regroupement d'ouvrages dont le nom de l'auteur commencerait par la lettre A. L'ensemble construit sur cette base semblerait

purement arbitraire et ne tiendrait compte ni des dates de publication, ni des genres textuels, ni des nationalités des auteurs, etc. De même, une masse d'exemples recueillis par le chercheur, sur la provenance desquels il ne pourrait fournir aucune indication telle que le support, la date de recueil, le contexte de production, etc. ne pourrait pas prétendre au titre de corpus par le caractère disparate et non contrôlé de cet assemblage composite. Un ensemble d'exemples recueillis à la volée peut donc difficilement avoir le statut de corpus :

> « Le recueil des exemples « à la volée » est souvent plus subjectif qu'on ne le souhaiterait : l'oreille sélectionne certaines choses et n'en entend pas d'autres » (Chanet, 2001 : 56)

Cela ne signifie nullement que ce genre de recueil n'est pas valide ou intéressant pour mener une étude en linguistique, pour peu que l'on dispose d'informations sur leur contexte de production. Il ne peut simplement pas recevoir l'appellation de corpus pour une partie des linguistes. Une stratégie possible est d'aller rechercher dans les corpus les plus diversifiés possibles les faits que les chercheurs ont relevés à la volée, en guise de contrôle.

La composition des corpus est étroitement liée aux objectifs que se donne le chercheur. Dans cette perspective, il arrive couramment que l'on distingue un corpus proprement dit (c'est-à-dire un ensemble structuré, choisi avec un objectif précis) et une banque de données (un ensemble de textes, généralement déjà constitué, ne correspondant pas précisément aux objectifs du chercheur). C'est souvent une différence de point de vue entre les concepteurs et les utilisateurs qui conduit à modifier / adapter la désignation de l'objet. De nos jours, de plus en plus de corpus sont utilisés avec un tout autre objectif que les objectifs initiaux, ce qui pose la question cruciale de la réutilisabilité des données. Par exemple, peut-on utiliser des données collectées pour documenter des phénomènes phonologiques comme la liaison ou le schwa pour mener à bien une étude à visée grammaticale ?

Sur cette question de l'utilisation de données constituées par d'autres, Reppen (2011) formule une remarque de bon sens à garder à l'esprit : il importe, avant de constituer son propre corpus, de s'assurer qu'il ne fasse pas doublon avec une ressource déjà disponible, précaution salutaire qui conduira le chercheur à expliciter les besoins auxquels doit répondre le corpus qu'il souhaite constituer. Il rappelle aussi des étapes chronologiques à même d'accompagner la réflexion : poser clairement l'objet de la recherche, s'interroger sur les droits d'utilisation et de diffusion des textes ou sur l'autorisation d'enregistrements des locuteurs, construire le corpus (inédit ou à partir de ressources existantes).

Sur la constitution elle-même, il convient d'être conscient que le corpus n'est pas un objet froid ou mort, qu'il résulte de choix (dont le chercheur doit être averti et conscient) et de décisions qu'il doit trancher à toutes les étapes. C'est pourquoi on conseille parfois aux chercheurs débutants de tenir un carnet de bord qui permet de noter les différentes décisions (par exemple les critères justifiant la sélection ou la mise à l'écart de textes) prises lors de la constitution et/ou de l'exploration du corpus. C'est un point important (et même essentiel) que souligne Condamines (2003 : 32) :

> « il semble difficile que l'on puisse constituer un corpus sans avoir une idée précise de ce que l'on veut en faire (hypothèse linguistique ou application) […] La constitution du

> corpus n'est pas indépendante de l'étude que l'on va mener et la réflexion sur les critères à mettre en œuvre doit être très élaborée. ».

Rastier (2004) est encore plus précis sur les incidences :

> « Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin, la structuration même du corpus. »

Dans cette notice, le terme corpus aura une valeur générique de collection de textes ou de transcriptions d'oral (accompagnées éventuellement de l'audio ou de la vidéo) utilisée ou recueillie pour mener un travail de description de la langue.

2. ÉMERGENCE DE LA LINGUISTIQUE DE CORPUS

2.1. La tradition francophone

Avant même que le terme de corpus ne s'impose à la communauté des linguistes, de nombreux auteurs de travaux, souvent de grande ampleur, ont eu le souci de rassembler des données en vue de les exploiter. Il nous a paru intéressant de regarder, avec un œil contemporain, ces tentatives et de mettre en lumière leurs visées et leur caractère précurseur.

Les *Archives de la parole* (1911-1914), réunies sous l'impulsion de Ferdinand Brunot, étaient destinées avant tout à conserver des grandes voix et des usages régionaux en voie d'extinction. L'objectif patrimonial est mis en avant. Cordereix (2014) parle de « bibliothèque sonore » pour caractériser cette vaste entreprise (à rapprocher d'un site comme celui de l'Institut National de l'Audiovisuel pour les archives radios et télés). Bergounioux (2014) relève le contenu disparate des enquêtes (qui récoltent notamment des « chansonnettes »), avant de conclure (p. 390) : « Ce n'est pas la fabrique d'un corpus mais un conservatoire d'espèces vouées à la disparition ». Les documents sonores enregistrés n'ayant pas connu d'exploitation linguistique d'ampleur, il les considère comme des « archives orales » mortes.

L'ouvrage de Jacques Damourette et Édouard Pichon (1911-1927) *Des mots à la pensée. Essai de Grammaire de la Langue Française* contient plus de 30.000 exemples dans 7 volumes (Muni Toke, 2012). L'une des originalités de leur approche tient aux données exploitées : écrites (majoritairement littéraires) mais aussi orales. Les auteurs des productions orales reproduites dans l'ouvrage sont recensés dans le glossaire qui précise leur date de naissance, leur situation sociale et leur « parlure » (bourgeoise ou vulgaire). Muni Toke (2013) décrypte l'idéologie qui oriente certains choix. Ce souci de documenter les caractéristiques des informateurs, s'inscrivant dans une vision de type sociologique, préfigure l'importance accordée de nos jours aux métadonnées (c'est-à-dire aux informations documentaires sur le corpus). Pour toutes ces raisons, Damourette et Pichon peuvent être considérés comme des précurseurs de la linguistique de corpus.

La grammaire des fautes d'Henri Frei (1929) s'appuie sur un ensemble de documents originaux : en majorité des lettres que des familles ont écrites à des soldats prisonniers, que Klinger (2011) nomme *corpus* ou *corpus épistolaire*. Cette collection, relativement homogène du point de vue des critères externes (il s'agit de disposer de productions de

scripteurs peu lettrés), permet à Frei de recenser ce qui est considéré par les discours prescriptifs comme des « fautes ». Dans la perspective de Frei, ces « fautes » manifestent des régularités qu'il est important de mettre en évidence.
Dans une perspective similaire, Henri Bauche (1920) fournit des précisions sur les données qu'il a collectées en vue de décrire le langage populaire :

> « J'ai simplement reproduit des phrases que j'ai entendues dans la rue, dans l'armée, dans les ateliers, les usines et les boutiques, chez les marchands de vin, dans les compartiments de troisième classe, dans les quartiers populaires de Paris et, aussi, des phrases que j'ai collectionnées dans des lettres écrites par des gens du peuple. » (p. 29)

Il semble donc s'appuyer sur une collection d'exemples authentiques dont la collecte reste malgré tout quelque peu opaque : les sources sont effacées du texte final.
Le contraste est frappant avec les deux ouvrages de Philippe Martinon (1913) et (1927). Cet auteur met en avant l'importance de l'usage en général et plus précisément de l'usage parisien, pour peu que celui-ci ait une certaine extension géographique : « La prononciation parisienne est la bonne, mais à condition qu'elle ne soit pas exclusivement parisienne, auquel cas elle devient simplement dialectale. » (1913 : VII). L'auteur ne donne pas beaucoup de précisions sur les informateurs qu'il a retenus et en dernier ressort il semble être le juge arbitre qui choisit « quelle est la prononciation qu'[il] tient en général pour la meilleure. » (*ibid*.).
On peut également citer les dialectologues et les sociolinguistes qui collectent des corpus et décrivent la langue en usage depuis bien longtemps. Dans ce domaine, l'invention du magnétophone puis de l'enregistrement vidéo a été une véritable révolution.

2.2. L'école de Londres

Si les descriptions basées sur des données authentiques existent depuis bien longtemps en linguistique, on fait généralement remonter l'invention de la linguistique de corpus en tant que discipline à John Rupert Firth et, à sa suite, John Sinclair. Firth est particulièrement connu pour sa célèbre phrase sur l'analyse distributionnelle (la description d'une unité linguistique basée sur ses contextes d'emploi) : « *you shall know a word by the company it keeps* » ; et Sinclair pour ses réflexions pratiques et méthodologiques détaillées. Le tableau ci-dessous, extrait de Léon (2008), synthétise les principales figures de l'école de Londres, à l'origine du développement de ce courant, ainsi que sa scission en deux approches dans les années 1990 : l'approche *corpus-driven* et l'approche *corpus-based*.

London School		
Fondateurs	Daniel Jones (1881-1967) et John Rupert Firth (1890-1960)	
2e génération	M.A.K. Halliday (né en 1925)	Randolph Quirk (né en 1920)
3e génération	John Sinclair (1933-2007) Chef de file du courant *corpus driven*	Geoffrey Leech (né en 1936) Chef de file du courant *corpus based*

Tableau 1. Les figures fondatrices de l'école de Londres (tiré de Léon, 2008)

L'approche dite *corpus-driven* repose sur le postulat que les corpus vont pouvoir faire émerger des hypothèses et que ces dernières n'ont pas besoin d'être formulées en amont de l'exploitation du corpus. C'est donc en observant les exemples dans des corpus que l'on pourra rendre compte du fonctionnement des langues. L'approche *corpus-based*, quant à elle, pose comme principe la formulation préalable d'hypothèses, leur vérification se faisant dans un second temps grâce aux données extraites de corpus.
Ces deux courants mettent en avant le caractère incontournable des exemples attestés pour décrire les langues et les problèmes que posent les exemples inventés. Par exemple, selon Sinclair (1991 : 6), « One does not study all of botany by making artificial flowers ». Dans cette citation, les fleurs artificielles représentent les exemples inventés, lesquels ne seraient pas adaptés à l'étude des langues. Nous reviendrons sur l'opposition entre exemples inventés et exemples authentiques un peu plus loin dans cette notice.
Une autre citation célèbre de John Sinclair est la suivante, traduite par Blanche-Benveniste (1996 : 25) :

> « Quand on invente des exemples, on confond souvent l'exemple et l'explication, l'exemple étant construit précisément pour justifier et clarifier l'explication. On ne peut pas inventer ce qu'est l'usage ; on peut seulement l'enregistrer. »

Cela met en lumière deux apports majeurs de la linguistique de corpus : la limitation du risque de circularité entre théorie et données et la nécessité d'observer des données non-sollicitées par l'analyste.
Dès les débuts de la linguistique de corpus informatisée, un outil s'est avéré incontournable : le concordancier. Il s'agit d'un logiciel permettant de retrouver l'ensemble des occurrences qui correspondent à une requête, accompagnées de leurs contextes droit et gauche. Cet outil est encore très largement utilisé de nos jours, simplement pour décrire la langue mais aussi dans des contextes d'enseignement. Un des premiers à avoir théorisé en profondeur l'usage des concordances est Sinclair (1991) (avec une mise à jour dans Sinclair, 2004). Les concordances permettent d'observer l'entourage d'une unité linguistique afin d'en préciser le fonctionnement. Cela a rendu possible la mise en œuvre systématique de l'approche structuraliste fondée sur l'analyse distributionnelle. John Sinclair a également popularisé la notion de collocation. On appelle *collocation* (terme forgé par J.R. Firth dans les années 1950) des éléments linguistiques qui entretiennent un rapport de sélection réciproque comme la relation qui existe entre *prendre* et *douche* (dans *prendre une douche*) ou *donner* et *conférence* (dans *donner une conférence*). On peut toutefois faire remonter à Bally (1909) et à ce qu'il désignait par « groupes phraséologiques » les réflexions autour des phénomènes de collocation [>Notice Legallois sur les constructions].
Sur ce point, on pourra également lire avec intérêt l'évocation par Halliday (1992) des débuts du travail de constitution de corpus en collaboration avec John Sinclair. Il défend dans ce texte l'intérêt des corpus (notamment de langue parlée) et le rapprochement entre lexique et grammaire que le corpus induit.

2.3. Corpus et enseignement des langues

Au départ, la linguistique de corpus répond prioritairement au souci d'enseigner l'anglais langue étrangère. Pour le français, la situation est identique avec des projets comme le

Français fondamental (Gougenheim *et al.*, 1964) dont l'objectif, au début des années 1950, était l'enseignement et la diffusion du français à partir de données écrites et orales. Le corpus *ESLO* (Enquêtes SocioLinguistiques à Orléans), longtemps désigné comme « le corpus d'Orléans », avait également cette visée-là, à savoir l'enseignement du français en Angleterre, les chercheurs à l'origine de cette ressource étant britanniques et non français (avec en plus un intérêt pour la dimension sociolinguistique). Depuis quelques années, on assiste à un retour massif vers les exploitations didactiques des corpus francophones à travers l'apprentissage sur corpus (Boulton et Tyne, 2014). On peut citer le projet *PFC* (Phonologie du Français Contemporain) et son prolongement en *PFC-EF* (Enseignement du Français), *CLAPI* (Corpus de LAngue Parlée en Interaction) et *CLAPI-FLE*, *OFROM* (corpus Oral de Français de Suisse Romande) et *OFROM-Enseignement* ou encore *ESLO-FLEU* (FLE et Linguistique pour l'Enseignement Universitaire), *FLEURON* (Français Langue Étrangère Universitaire Ressources et Outils Numériques) et *FLORALE* (Français Langue Orale pour le FLE).
La langue la mieux dotée, tant du point de vue des méthodes pour l'enseigner que des corpus constitués, est sans conteste l'anglais. On peut citer par exemple le *Brown Corpus* dans les années 1960 composé de 500 échantillons de 2000 mots ou le *London-Lund Corpus* pour l'anglais parlé. Mais le corpus qui a sans doute eu le plus d'impact à ce jour est le *British National Corpus* (100 millions de mots) qui a permis entre autres la réalisation de la première grammaire d'une langue sur corpus écrits et oraux, à savoir la *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999) avec des déclinaisons pour les apprenants.

2.4. Un domaine encore mouvant

Comme nous l'avons vu, il n'existe pas de consensus sur la définition du corpus. Il n'en existe pas non plus sur la manière de les utiliser (si l'on dépasse la simple idée de travailler à partir de données authentiques). Par exemple, en témoigne la différence que nous avons soulignée entre *corpus-based* et *corpus-driven* (voir section 2.2.). Et même à l'intérieur de l'approche *corpus-based*, on observe des pratiques bien différentes. L'utilisation du corpus en tant que réservoir d'exemples (dont on ne retient que ce qui va dans le sens de l'analyse) ou bien, au contraire, une prise en compte plus fouillée, plus complexe. L'exemple de la *Grande Grammaire du Français* (voir la présentation générale dans Cappeau, 2021) se situe dans cette perspective illustrative en proposant de surcroit des exemples oraux relus, ce qui dénature quelque peu le corpus originel. Cette utilisation du corpus comme base d'exemplification est à l'image de la *Cambridge Grammar of the English Language*. C'est une tout autre perspective que celle choisie par la grammaire de Biber *et al.* (1999) pour l'anglais, qui prend le corpus comme point de départ de l'analyse.
Il existerait également, selon Williams (2006), une distinction à faire entre deux linguistiques, que l'on peut faire ressortir en français par l'intermédiaire du choix de la préposition : linguistique sur corpus et linguistique de corpus. Il s'agit de distinguer, d'un côté, une discipline à part entière, autonome, ayant des exigences fortes quant à sa méthodologie et ses choix épistémologiques : la linguistique *de* corpus. Et, d'un autre côté, la linguistique *sur* corpus, qui prendrait simplement le corpus comme support de l'analyse sans base épistémologique précise. Cette seconde acception est à l'interface

avec d'autres disciplines et servirait à d'autres applications que la seule linguistique (sociologie, traitement automatique des langues, etc.).
À l'heure actuelle, différentes pratiques coexistent, allant des plus artisanales aux plus technologiques, pratiques que nous détaillons dans la section 5. Pour l'instant, nous allons nous intéresser à l'opposition qui a longtemps perduré entre linguistique introspective et linguistique de corpus.

3. OPPOSITION DE DEUX LINGUISTIQUES

En linguistique, l'intérêt de travailler à partir de corpus n'a pas été d'emblée validé par l'ensemble de la communauté, ce qui a donné lieu à une fracture – longtemps violente – entre deux linguistiques.

3.1. Linguistique introspective et linguistique de corpus

L'inscription dans une approche sur corpus induit une position forte par rapport à la place et à la nature des données langagières prises comme matériau principal ou exclusif de la description. En effet, l'objectif de la linguistique de corpus est de décrire des faits, des observables attestés à partir de vastes collections de textes authentiques de différentes provenances. Ce positionnement épistémologique fort a donné lieu à de nombreuses critiques et a mis du temps à être perçu comme légitime dans une discipline se voulant scientifique. Et il arrive encore au XXI^{ème} siècle que cette méthodologie soit critiquée (Scheer, 2004 ; 2013).
L'approche sur corpus a dû se faire une place face à la grammaire générative de Noam Chomsky et à des approches structuralistes telles que le lexique-grammaire de Maurice Gross. Les tables du lexique-grammaire répertorient un grand nombre de contraintes syntaxiques, influençant le fonctionnement des unités lexicales, à partir principalement de l'intuition du locuteur natif. Les données authentiques apparaissaient insuffisantes à couvrir les possibles linguistiques. L'approche générative est centrée sur ce qui est possible (la compétence du locuteur natif à partir de l'introspection et des jugements de grammaticalité) alors que ce qui fonde la linguistique de corpus, c'est la description de ce qui est attesté.
La critique la plus virulente contre la linguistique de corpus est venue de Noam Chomsky pour qui elle n'existe tout simplement pas (propos tenus dans le cadre d'un entretien et rapportés par Aarts, 2000 : 5). Selon Chomsky, les exemples attestés ne sont pas de bons candidats en vue de la modélisation des langues. Il s'inspire de la dualité saussurienne langue/parole – qu'il rebaptise compétence/performance – et désigne la compétence comme étant l'objet exclusif de la linguistique. Dans ce schéma, la performance – impure par nature – ne peut donner accès au système. Chomsky s'intéressant à la grammaire interne des locuteurs (et donc à la compétence), les corpus ne permettraient pas de l'observer et seraient donc disqualifiés à ses yeux. La performance serait par essence perturbée par la situation dans laquelle prend place le dialogue ou le monologue : la fatigue, une pathologie éventuelle… Les données authentiques seraient donc fortement bruitées.
Noam Chomsky s'appuie en réalité sur ce qui se fait dans d'autres disciplines scientifiques (notamment dans les sciences dites « dures »), dans lesquelles il est courant

de neutraliser des paramètres pour pouvoir modéliser des phénomènes, à l'image des cours de physique dans lesquels on neutralise le frottement d'un mobile glissant sur un plan incliné. Cette approche est tout à fait justifiée quand l'observation directe n'est pas possible. Mais en linguistique, une telle démarche peut amener à neutraliser la variation, caractéristique incontournable des langues naturelles.
Autre critique que l'on entend régulièrement : le caractère par essence fini du corpus face au nombre théoriquement infini d'énoncés. Cela peut aller jusqu'à l'affirmation qu'il faut abandonner le corpus pour le remplacer par de l'introspection, plus à même de générer l'ensemble des énoncés possibles (Sctrick, 1968).
On le comprend ici, la conception générative de la grammaire défendue par Noam Chomsky, d'un côté, et les approches sur corpus, de l'autre côté, représentent deux perspectives antagonistes de la linguistique. Cela touche à la conception profonde de la discipline et de ses objectifs. Dans un cas, il s'agit d'élaborer des modèles génératifs de fonctionnement de la langue (ou des langues) en envisageant ce qui est possible et impossible. Pour ce faire, on recourt à « l'intuition du locuteur natif » et donc à des jugements de grammaticalité et à des exemples inventés grâce à l'introspection, ce qui peut poser problème à cause notamment des représentations des locuteurs et de l'influence des discours prescriptifs. Cette manière d'analyser la langue comporte donc des biais qui font écho à la critique de Chomsky au sujet des corpus : il s'agit là aussi de données bruitées. Dans la seconde perspective, l'objectif est d'établir les régularités observées dans les productions effectives des locuteurs.
Cependant, ce débat, tel qu'il a été posé dès les années 1960, a perdu de sa pertinence au XXI^{ème} siècle tant la linguistique de corpus s'est développée dans le paysage de la science linguistique contemporaine. Il faut dire que les premiers corpus étaient d'une taille réduite, peu accessibles et difficiles à exploiter, ce qui peut expliquer en partie les critiques initiales. Mais ce n'est plus le cas. De nos jours, les corpus ont changé de dimension, ils sont nettement plus volumineux et représentatifs et leur usage s'est largement démocratisé. Désormais, à la suite notamment de Jacques (2005), on pourrait dire qu'il s'agit d'un choix entre linguistique introspective et linguistique de corpus, plutôt qu'une opposition ou une confrontation.

3.2. Un rapprochement possible

Comme nous venons de le voir, la linguistique de corpus et la linguistique introspective ont vécu chacune dans une relative autonomie pendant plusieurs décennies, avec ponctuellement des critiques virulentes de part et d'autre. Mais depuis le début des années 2000, la situation a quelque peu évolué. Comme l'explique Teubert (2009) :

> « La linguistique de corpus n'entend pas se poser en alternative ou en concurrente face aux paradigmes qui prétendent découvrir, ou du moins modéliser, la réalité d'une faculté de langage spécifique ou d'une faculté universelle de langage. »

L'opposition entre les deux linguistiques est désormais à reconsidérer. En réalité, les deux linguistiques ne travaillent pas véritablement sur le même objet. La linguistique introspective donne accès aux représentations des locuteurs (un peu comme les sondages) et aux tournures potentielles employées dans une langue donnée alors que la linguistique de corpus permet d'observer ce qui est réellement en usage. Par exemple, pour une

tournure, une forme graphique ou une prononciation, on peut mesurer l'influence des injonctions normatives sur les réponses des locuteurs : pense-t-on que l'on emploie *malgré que* ? Prononce-t-on *consensus* avec [ã] ou [ɛ̃] ?

La connaissance de la forme correcte influence notre jugement et pas forcément nos usages ou dans des proportions qui peuvent être moindre. Dit autrement, on peut rejeter une tournure que l'on emploie pourtant régulièrement. La science participative (ou crowdsourcing) via des plateformes sur internet permet de collecter en un temps réduit une quantité de jugements sans commune mesure avec les techniques d'enquête antérieures (Avanzi *et al.*, 2016). La linguistique de corpus, quant à elle, aborde la question de l'attesté, de ce qui est réellement utilisé par les locuteurs.

Il est toutefois possible de concilier ces deux approches car cela permet d'aborder les faits de langue à partir de deux points de vue différents. En effet, des tournures jugées grammaticales ou acceptables peuvent être absentes des corpus consultés. La linguistique introspective permet de mesurer des degrés d'acceptabilité là où la linguistique de corpus mettra plus en avant la fréquence ou la présence/absence.

On a également assisté, ces dernières années, à l'essor de la linguistique expérimentale dans le prolongement de la linguistique introspective. La linguistique expérimentale a pour objectif principal de modéliser les phénomènes linguistiques à partir de données collectées dans des conditions contrôlées. Le principe est d'isoler des paramètres exerçant une influence potentielle et d'observer dans quelle mesure cela affecte les productions des locuteurs. Par exemple, en quoi le nombre de syllabes d'un adjectif épithète permet-il de prédire sa position par rapport au nom dont il dépend ? Les corpus permettent difficilement ce genre de travaux étant donné la faible probabilité de trouver un grand nombre d'exemples identiques à l'exception d'un seul trait distinctif.

Des travaux actuels, en syntaxe notamment, combinent linguistique de corpus et linguistique expérimentale pour tenter de cerner certains phénomènes linguistiques comme la place de l'adjectif épithète ou l'ordre des compléments (Thuilier, 2012) ou bien encore l'acquisition des interrogatives (Thiberge, 2020). Ces travaux abordent ces questions en termes de contraintes préférentielles plutôt que catégoriques et utilisent généralement des méthodes statistiques avancées comme la régression logistique. Cela permet de mesurer l'incidence d'un ensemble de paramètres sur les variations observées dans les productions d'un panel de locuteurs mis en condition de produire la structure objet de l'analyse.

Cependant, il est indispensable de tenir compte de l'influence de la composante prescriptive sur les jugements et de la catégorie de la population qui sert de « cobaye » à ces expérimentations. Ce sont généralement des étudiants ou des personnes ayant une bonne maitrise du français écrit qui jugent en grande partie en fonction de paramètres normatifs et qui ont une conscience souvent limitée de leurs usages réels. On pourrait formuler la même critique à l'encontre des corpus, dans lesquels la catégorie des locuteurs éduqués est surreprésentée. Mais une recherche de situations diversifiées de parole spontanée permet de limiter la probabilité d'une surreprésentation des tournures normées. Certains projets donnent justement à voir des productions écrites de peu lettrés (*Corpus 14*, *Prize papers*), et à entendre des productions orales de locuteurs de provenances diverses (*MPF*).

4. CARACTÉRISTIQUES DES CORPUS

Ce chapitre porte sur les caractéristiques des corpus et aborde la plupart des paramètres incontournables à interroger lorsque l'on souhaite en constituer un ou se servir d'un corpus déjà existant.

4.1. Taille

Les corpus sont de taille extrêmement variable et leur volume n'a cessé de croitre. Dès les années 1980, *Frantext* et ses 160 millions de mots étaient interrogeables à distance moyennant la création d'un compte ou bien par l'intermédiaire de stations d'interrogation se trouvant dans des bibliothèques parisiennes. En juin 2024, *Frantext* comporte 5 679 ouvrages pour 272 millions de mots et est accessible dans le monde entier via un abonnement. Mais ce corpus a, pendant longtemps, fait figure d'exception, les données informatisées pour le français, en dehors de *Frantext*, ayant des tailles nettement plus modestes jusqu'au tournant des années 2010 où le nombre et la taille des corpus disponibles se sont fortement accrus.
Il existe également un déficit du côté de l'oral par rapport à l'écrit. Ce décalage entre oral et écrit concerne toutes les langues et pas seulement le français, même si on peut supposer que la tradition littéraire du français a représenté un frein important concernant les incitations institutionnelles à constituer des corpus oraux.
À l'heure actuelle, la taille des corpus varie selon de nombreux facteurs d'ordre technique ou de représentation/valorisation sociale. On en cite quelques-uns :

• Le médium : du côté de l'écrit, il est désormais assez facile de collecter des données massives grâce au web, mais il peut être difficile d'identifier les locuteurs à l'origine des textes récupérés (problème de traçabilité) et de les diffuser à cause de questions juridiques. De plus, il est presque impossible de s'assurer de manière systématique que les textes récupérés n'émanent pas d'un locuteur non-natif, d'une traduction, automatique ou pas, ou qu'ils n'ont pas été élaborés par une intelligence artificielle générative. Les données orales, quant à elles, nécessitent un temps important pour leur collecte et leur transcription. Ainsi, il peut être tentant de recourir à des sous-titres de films ou de séries. Mais cette pratique pose question pour l'étude du français parlé. Tout d'abord, les sous-titres ne reproduisent pas *in extenso* ce qui est dit. Ensuite, ces données ne sont pas de même nature que le français non-planifié étant donné qu'elles émanent d'un support écrit oralisé (des dialogues reconstitués). Il s'agit au mieux, dans ce cas, d'oral représenté. Des travaux de recherche portent spécifiquement sur des données représentant des dialogues oraux, particulièrement en diachronie à partir de textes littéraires et de pièces de théâtre (Marchello-Nizia, 2014).

• Le genre des productions collectées : pour les écrits, la littérature et la presse nécessitent des accords avec les maisons d'édition et les journaux. Les copies d'élèves requièrent, comme les transcriptions de français parlé, un long travail de collecte et de saisie du texte (Doquet *et al.*, 2017). Pour l'oral, les interactions dans un commerce sont moins faciles à collecter que les entretiens. Les consultations dans un cabinet médical sont, quant à elles, impossibles à diffuser pour des raisons de confidentialité.

• Le sujet d'étude : si l'on souhaite travailler sur les productions d'un auteur en particulier, il va de soi que le corpus est dépendant du nombre d'ouvrages publiés par cet auteur. De même en français parlé, le fait de vouloir travailler sur un nombre limité de locuteurs à des âges différents (corpus longitudinal) ou bien sur des locuteurs différents (corpus transversal) va forcément avoir une incidence sur le volume de données exploitables.

Par conséquent, on le voit, ces paramètres vont induire la mise à disposition ou la constitution de corpus de taille très variable. Pour l'écrit, la taille de la plupart des corpus se situe entre la dizaine de millions et le milliard de mots. Certains corpus vont même jusqu'à 10 milliards de mots. La plus grande banque de données existant à l'heure actuelle est sans doute celle de *Google Livres*, contenant l'ensemble des ouvrages numérisés par Google. Mais celle-ci n'est pas disponible pour les linguistes et le grand public, en dehors de l'outil *Google ngram* qui permet de générer des graphiques chronologiques pour observer l'évolution de la fréquence du motif recherché (généralement un mot ou un groupe de mots). On a principalement accès à ce graphique chronologique, ce qui rend les résultats impossibles à contrôler et donc soumis à caution. D'autant que la plupart des ouvrages ont fait l'objet d'une reconnaissance automatique de caractères et comportent donc des erreurs.
Une autre ressource incluant un ensemble de données textuelles très volumineuses est hébergée par le site *Gallica* de la Bibliothèque nationale de France. Il y a notamment une grande quantité d'articles de presse qui ont été numérisés ainsi que de nombreux ouvrages en langue française. Tous ces corpus (et d'autres) sont accessibles via le portail *Gallicagram* (https://shiny.ens-paris-saclay.fr/app/gallicagram).
Pour l'oral, les corpus actuels atteignent bien souvent plusieurs centaines de milliers de mots (*PFC, CRFP, C-ORAL-ROM* partie française), parfois un peu plus d'un million de mots (*CFPP, OFROM, TCOF*) voire dépassent largement ce seuil (*CEFC* partie orale, *ESLO*).
La taille des données va avoir une incidence sur les champs d'étude possibles. Un très grand corpus pourra difficilement être exploré de manière systématique en vue de décrire un phénomène linguistique de forte fréquence et inversement un petit corpus ne permettra pas d'envisager des travaux sur les collocations. Pour le lexique, on sait qu'il est nécessaire de disposer de corpus très volumineux (plusieurs dizaines voire centaines de millions de mots au minimum). Par exemple, on ne trouve que 44 occurrences d'un lexème aussi banal que *corbeille* dans le *CEFC* (corpus de plus de 10 millions de mots). Certains phénomènes, y compris syntaxiques, pourront être très peu fréquents même dans des corpus de grande taille [>Notice adjectif adverbal]. Cela amène à se poser la question du nombre minimum d'attestations pour pouvoir envisager de mener à bien une étude descriptive (voir section 5.6.4.).
Il y a une tendance parfois à croire qu'il faut viser à tout prix les corpus les plus volumineux possible, idée qu'Habert (2000) a mise à mal. En effet, des corpus de grande taille ne sont pas adaptés à tout type d'études, même si cela peut permettre de fédérer une communauté (à l'image du *British National Corpus*) lorsque le corpus a une ambition de représentativité. Dans le cas du *BNC* justement, il y a eu un travail minutieux de sélection et d'équilibre et non uniquement de ce qui était facilement récupérable.

Avec la généralisation des techniques d'apprentissage profond pour alimenter des logiciels d'intelligence artificielle générative, la question du volume des données est revenue dans l'actualité linguistique. En effet, il est indispensable de disposer d'énormes quantités de données langagières (de l'ordre de plusieurs milliards de mots) pour entrainer les algorithmes. Les modèles de langue ainsi construits le sont essentiellement à partir d'écrits normatifs. Dans ce contexte, les corpus de langue spontanée de quelques millions de mots peuvent être considérés comme négligeables. Et ces outils permettent à leur tour d'analyser automatiquement de gros corpus impossibles à traiter manuellement. Cela comporte à la fois des bénéfices et des risques : d'un côté, la fouille et le tri automatiques de masses de données trop volumineuses pour une analyse humaine, d'un autre côté, des traitements dont la qualité est difficile à évaluer.
Si les gros corpus peuvent convenir pour certaines tâches, il ne faut pas perdre de vue qu'une grande quantité de données fortement bruitées (contenant des doublons, des traductions automatiques, etc.) ne peut aboutir qu'à des descriptions dont la qualité/fiabilité est très incertaine.
Pour conclure, comme le résume très bien Desagulier (2017 : 6) :

> « La petite taille devient un problème si l'unité qui vous intéresse n'est pas bien représentée. En résumé, la taille est importante, mais si elle est utilisée à bon escient, un petit corpus vaut mieux qu'un grand corpus utilisé de manière inconsidérée. » (trad. Cappeau)

4.2. Exhaustivité

Pendant des siècles, le terme *corpus* a été utilisé pour désigner des ressources documentaires exhaustives. Cela correspondait à tous les documents disponibles sur un domaine donné (Bilger éd., 2000b : 11). Ainsi, pour connaitre un sujet, il suffisait de consulter le corpus. La couverture d'un corpus est toujours un sujet de préoccupation, mais il va de soi que dans ce domaine l'exhaustivité est un idéal jamais atteint (en dehors de cas très spécifiques). En effet, comment pourrait-on constituer un corpus contenant l'ensemble des énoncés produits dans une langue ou même dans un genre donné comme la presse écrite par exemple ? La quête d'exhaustivité est certainement en lien avec la taille importante que possèdent de nombreux corpus. C'est par exemple une association que l'on trouve dans Mayaffre (2010b : 45) :

> « Nous avons en effet, du mémoire de Maîtrise en 1995 à la thèse que nous co-dirigeons aujourd'hui, cherché à appréhender de gros corpus, souvent exhaustifs […] ».

Le terme *exhaustif* est souvent délimité par des expressions comme *si possible* (ou *as it can be*), ce qui en atténue la portée.
La question de l'exhaustivité se pose différemment selon les données qui sont rassemblées. Mayaffre y recourt à plusieurs reprises : « Lorsqu'un corpus – exhaustif si possible – arrive à témoigner de lui-même, ne lui demandons pas de témoigner, en plus, d'autres réalités. » (2010b : 43). Il est vrai que son domaine d'étude privilégié rend atteignable cette perspective : « Il n'existe en effet plus de difficulté technique à recueillir et à traiter l'ensemble du dictionnaire ; les corpus lexicographiques peuvent donc non seulement être des corpus clos mais des corpus finis. » (Mayaffre, 2005). D'autres travaux qui prennent appui sur des dictionnaires peuvent également viser l'exhaustivité :

par exemple, Duchet *et al*. (2012) ont rassemblé toutes les indications morphologiques présentes dans des dictionnaires anglais d'une période précise. Les corpus clos et des périodes bien délimitées sont des facteurs qui permettent d'atteindre l'exhaustivité des données recueillies.

Dans d'autres domaines comme les corpus de textes, « on ne peut prétendre rassembler tous les énoncés possibles » (Vaguer, 2007 : 209). Même lorsque les conditions semblent réunies pour atteindre l'exhaustivité, Leech (2015) soulève des réserves que devrait garder en mémoire tout chercheur. Il prend l'exemple de langues mortes, pour lesquelles on peut considérer disposer de données dans leur intégralité, dans l'attente de nouvelles découvertes. Comme il le précise :

> « les corpus de données qui nous sont parvenus sont le résultat de survivances fortuites, ne contiennent bien sûr aucune langue parlée et sont généralement fortement biaisés en faveur de certaines périodes, genres et auteurs. » (Leech, 2015 : 147 ; trad. Cappeau).

Ce n'est donc, au mieux, qu'une exhaustivité partielle ou bancale qui est obtenue. De façon comparable, Mellet pointe les limites et les réserves associées à cette notion :

> « Un corpus ne peut être ***clos et exhaustif*** que dans le cadre d'une monographie, auquel cas il sera étudié en tant que tel, sans prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation. » (Mellet, 2002 : 6).

Finalement, s'interroger sur l'exhaustivité des données collectées présente un mérite : obliger à se poser la question de la délimitation la plus précise possible du corpus. Si l'on prend l'exemple d'Alain-Fournier, on peut considérer que son roman *Le Grand Meaulnes* constitue, à lui seul, la totalité de sa production romanesque publiée. Mais faut-il alors inclure les variantes du texte que présente l'édition de la Pléiade ? Si l'on veut s'appuyer sur l'ensemble de sa production publiée, il faudra aussi inclure *Miracles* (une série de textes courts parus dans divers journaux). Pour approcher au plus près l'usage particulier de la langue écrite de cet auteur, doit-on aussi intégrer sa correspondance ? Même en prenant en compte tous ces éléments, il faudra considérer que ce n'est qu'une partie de ses écrits qui sont ici rassemblés : il manquera toujours des lettres perdues, des notes non conservées, des écrits éphémères. Les problèmes seront encore plus complexes dès lors que l'objet d'étude sera plus vaste (par exemple la langue littéraire autour de la guerre de 1914-1918).

De fait, pour de nombreux sujets d'étude, la notion d'exhaustivité n'a pas grand sens : on ne peut pas imaginer recueillir tous les écrits d'une période donnée, tous les échanges oraux d'un individu (on peut l'envisager pour les enfants jeunes dont les interactions sont en nombre réduit, pendant un temps court, en les enregistrant).

Selon les disciplines, l'exhaustivité n'est pas un objectif déterminant, comme le précise Maingueneau (2005 : 73) :

> « Les analystes du discours peuvent également construire des corpus d'éléments de divers ordres (lexicaux, propositionnels, fragments de textes) extraits de l'interdiscours, sans chercher à construire des espaces de cohérence, à constituer des totalités. »

Enfin, dans la phase d'exploitation des données, on retrouve parfois la notion d'exhaustivité qui signifie alors que le chercheur doit récupérer la totalité des exemples

disponibles dans le corpus afin d'en proposer une analyse : il s'agit là d'une précaution méthodologique qui prémunit le chercheur contre le risque d'écarter des exemples avant analyse ou d'opérer une sélection (sans en expliciter les critères) qui pourrait fausser la description.

4.3. Représentativité

Un corpus, quelle que soit sa taille, ne contient qu'une portion de l'usage de la langue à une période, chez un ensemble limité d'individus, pour un type restreint de productions. C'est pourquoi on peut considérer, dans un premier temps, qu'un corpus ne représente que lui-même (« Par définition, tout corpus est seulement parfaitement représentatif de lui-même » (Aston, 2011 : 3 ; trad. Cappeau). Mais très souvent, les chercheurs qui ont constitué un corpus ont l'ambition que sa description ait une portée plus générale. Cette démarche n'a, en soi, rien d'inédit et rappelle les notions de modèle utilisé entre autres dans les sciences de la nature (Schmidt-Lainé et Pavé, 2008) : le travail sur un fragment, un modèle réduit peut être transposé à un ensemble bien plus vaste. Restent des questionnements à garder en tête, qui sont listés ci-dessous.
Tout corpus est lié à un objectif de recherche qui en délimite, en partie, le contour. Cela a une incidence sur la nature des faits collectés, de nombreux biais pouvant intervenir sur la sélection des données et rendre celles-ci plus ou moins représentatives et plus ou moins exploitables. Leech (2015) s'interroge ainsi sur les corpus de presse : quels critères de sélection privilégier ?

• Le lectorat ? (mais aucun échantillonnage sur la base du lectorat n'a été tenté, à notre connaissance)
• Les chiffres de diffusion ? (mais ils sont souvent difficiles à obtenir ou à vérifier)
• Le prestige culturel de la publication ?

> « Il est vrai que, dans le passé, certains concepteurs de corpus ont introduit des critères d'évaluation, jugeant, par exemple, qu'un journal à grand tirage (au Royaume-Uni) est plus important ou plus influent qu'un journal à sensation ; qu'un roman qui remporte un prix national de littérature est plus important qu'un best-seller de pulp-fiction ; ou que des locuteurs appartenant aux groupes socio-économiques A et B sont plus dignes de figurer dans un corpus que des membres des groupes socio-économiques inférieurs D et E. Cependant, cet élitisme est totalement déplacé dans un corpus destiné à l'analyse linguistique. » (Leech, 2015 : 140 ; trad. Cappeau)

Ces observations soulignent que le linguiste doit être attentif aux biais que pourraient introduire, dans son étude, ses propres préjugés relatifs à la légitimité sociale des documents qu'il utilise (ou au contraire rejette). On peut par exemple vérifier qu'au début des années 2000, peu de corpus écrits de grande ampleur étaient facilement accessibles en dehors de *Frantext* et d'une partie des archives du journal *Le Monde*. Cela a donné lieu à de nombreux travaux qui, pour la langue écrite, privilégiaient soit l'une soit l'autre ressource (Aliquot-Suengas, 2003 ; Gary-Prieur, 2005). Heureusement, ce n'est plus le cas aujourd'hui avec notamment la mise à disposition d'échanges épistolaires durant la première guerre mondiale (*Corpus 14*), divers corpus issus du Web ou de SMS (*88milSMS*).

La question de la représentativité gagne à être scindée en deux, selon l'objectif visé lorsque le corpus est constitué :

> « Il existe de nombreux types de corpus différents. Certains tentent d'être représentatifs d'une langue dans son ensemble et sont appelés « corpus généraux » ou « corpus de référence », tandis que d'autres tentent de représenter un type particulier d'utilisation de la langue et sont appelés « corpus spécialisés ». » (Cheng, 2012 : 4 ; trad. Cappeau)

L'idée sous-jacente à la collecte d'un matériau linguistique de très grande ampleur est de permettre l'élaboration de grammaires et de descriptions diversifiées (par exemple la grammaire de Biber *et al*., 1999 déjà citée). Dans ce cas, le corpus a pour ambition de fournir une représentation, sous forme de modèle réduit, de la langue elle-même. Biber (1993) a détaillé une démarche pour constituer, au mieux, un corpus répondant à ce cahier des charges. Plusieurs étapes doivent s'enchainer : le projet, le regroupement des textes, un examen plus précis de la variation afin d'adapter la conception initiale du corpus. Vaste programme mais comme le précise Leech (2015 : 134 ; trad. Cappeau) « aucune tentative (à ma connaissance) n'a été faite pour mettre en œuvre le plan de Biber pour construire un corpus représentatif ». En conséquence, les auteurs sont conscients des limites auxquelles ils sont confrontés, c'est pourquoi l'on trouve souvent des formules comme « jugé représentatif » ou « considéré comme représentatif » pour qualifier le corpus.
Dans le cas de corpus de moindre taille, « Définir un corpus oblige à circonscrire un objet clos et donc à se donner des règles pour garantir un minimum de représentativité à cet objet, cette représentativité étant elle-même liée à un objectif d'étude. » (Condamines et Dehaut, 2011 : 268). Décrire la langue utilisée dans le corpus peut avoir une portée plus limitée comme une meilleure couverture des dictionnaires :

> « l'objectif est d'accorder une attention spéciale à la langue des conversations familières, en vue notamment de la voir un jour mieux représentée dans les dictionnaires » (Dostie, 2016).

Pour finir, la notion de représentativité ne s'applique pas toujours au corpus lui-même. Les promoteurs contemporains des *ESLO* font ainsi la distinction entre le corpus et les pratiques linguistiques qu'il peut permettre d'étudier :

> « L'objectif du projet n'est pas de produire un corpus représentatif, mais d'offrir un réservoir de corpus conçu dans un souci de représentativité des pratiques linguistiques d'une communauté d'auditeurs dans une ville donnée, à des moments distincts. » (Baude et Dugua, 2016).

4.4. Diversité

La diversité dans les ressources utilisées est censée permettre d'observer de la variation dans les usages. La notion de diversité est plutôt utilisée par les chercheurs français alors que les auteurs anglo-saxons privilégient celle de corpus équilibré (*balanced*) en lien avec la notion de représentativité (voir 4.3.) :

> « La langue contenue dans ces corpus était principalement équilibrée de manière à constituer un échantillon représentatif de la langue dans son ensemble. » (Baker, 2008 : 146 ; trad. Cappeau)

> « Il existe une règle empirique à laquelle peu de gens sont susceptibles de déroger. En général, plus un corpus est important et plus il est diversifié en termes de genres et d'autres variétés linguistiques, plus il sera équilibré et représentatif. » (Leech, 2015 ; trad. Cappeau)

Pendant longtemps, le *BNC* a fait figure de modèle pour les corpus équilibrés. Il est composé de 10 % d'oral et 90 % d'écrit. Toutefois, cette répartition interroge quant à la place dévolue à ces deux médiums étant donné que cela ne représente nullement une proportion réaliste, ni en production, ni en réception pour la quasi-totalité des locuteurs existants. Si l'on creuse un peu, des contraintes d'ordre pratique plus que scientifique jouent un rôle important dans la répartition : par exemple, il est plus chronophage et donc couteux de collecter un million de mots produits à l'oral qu'un million de mots écrits (en dehors de cas limités comme les copies d'élèves) et la représentation sociale de la diversité des documents écrits est peut-être mieux connue et plus facile d'accès que celle des documents oraux.

L'intérêt de constituer ou d'utiliser un corpus diversifié dépend du projet de recherche envisagé. L'idée de corpus diversifié est importante lorsque l'objectif du chercheur est de ne pas limiter son analyse à un usage singulier mais de chercher à approcher des réalisations différenciées selon les contextes de production. Les deux paramètres (*diversité* et *représentativité*) ne convergent pas forcément : si l'on constitue un corpus de 500 000 mots de français oral de conversation, 500 000 mots de sermons, 500 000 mots de journal télévisé et 500 000 mots de littérature contemporaine, on pourra considérer, dans une certaine mesure, que l'on dispose d'usages contrastés et que le corpus est diversifié. Pour autant, cet assemblage est-il représentatif du français ? Des productions comme les sermons ou la présentation d'un journal télévisé ne relèvent pas de pratiques communes et sont donc peu représentatives voire marginales. Ainsi, de grands pans des pratiques linguistiques restent inaccessibles à l'étude. Cela soulève la question du poids que chacune de ces pratiques devrait avoir (voir 4.3.).

Le terme *diversité* s'applique aux textes mais est souvent étendu à d'autres aspects des documents rassemblés dans le corpus :

> « Ces corpus [=Les grands corpus de données orales authentiques] augmentent l'accessibilité de données quantitativement importantes et sociolinguistiquement diversifiées, en ouvrant ainsi de nouvelles possibilités pour la comparaison, l'analyse de la variation, le traitement qualitatif/quantitatif des formes linguistiques (cf. Aijmer & Altenberg, 1991 ; Kallmeyer, 1997 ; Bilger, 2000). » (Mondada, 2001 : 143)

> « Pour constituer le corpus MPF, on a renoncé à la problématique du vernaculaire théoriquement abordée dans l'opposition entre situations formelles et informelles (Labov, 1972) pour privilégier la qualité des interactions entre les protagonistes, qui correspondait mieux à nos hypothèses. Les options sous-jacentes à ce choix concernent les facteurs susceptibles de produire de la variation et de la diversification langagières dans une même langue : les facteurs décisifs relèvent-ils des situations ? Des genres ? Des caractéristiques sociodémographiques des locuteurs ? Des interactions ? » (Gadet et Guerin, 2016)

Dans la présentation du *Corpus d'Étude du français contemporain* (*CEFC*, Debaisieux et Benzitoun, 2020), la notion de diversité concerne l'origine des textes rassemblés (ils proviennent, entre autres, de plusieurs titres de presse distincts), aux usages du français

(conformément à la position de Blanche-Benveniste et Jeanjean, 1987 qui observaient une mise à l'écart d'une partie des usages), aux types de textes et aux locuteurs.
Habert (2000) fournit un inventaire des usages bien documentés (le français d'hier, le français écrit et le « bon usage ») et des usages sous-représentés (le français non-hexagonal, le français scientifique, le français non-standard). Les lignes ont bougé depuis cette date, mais il est bon de conserver l'idée de diversité et de ne pas limiter le français à un seul usage ou à un seul médium (si le projet de recherche s'y prête, bien évidemment).
Tous les chercheurs, selon leurs centres d'intérêt, ne mettent pas les mêmes choses sous le terme de corpus variés. Voici un exemple un peu ancien d'une « variété » malgré tout assez peu diversifiée : Vergne et Giguet (1998 : 30) qualifient de « corpus variés » l'ensemble constitué d'articles du *Monde*, de textes littéraires et de textes scientifiques. Seuls trois types d'écrits, assez fortement codifiés, sont retenus.
En syntaxe, il peut être intéressant de s'appuyer sur de l'écrit littéraire (mais pas uniquement) et de contraster avec de l'oral (mais pas seulement des conversations). La diversité des données peut permettre de cartographier les usages d'une tournure à une époque donnée. Des projets comme Orféo ont été en partie construits pour développer cette piste. En effet, les approches sociolinguistiques ont montré que les locuteurs utilisent des formes différentes (dans l'articulation, la prosodie, le lexique, les tournures syntaxiques, etc.) selon les contextes dans lesquels ils s'expriment par oral ou par écrit [> voir le point 1.3. de la notice *Langue et variation*]. Cela ressort aussi d'une certaine tradition des études stylistiques en littérature ou sur les genres littéraires.
Cependant, une partie des travaux en linguistique française n'aborde pas la question de la variation à partir d'une étude contrastive de données hétérogènes. Il arrive que les descriptions portent sur des corpus relativement homogènes (notamment *Frantext* pour la littérature ou le *French Treebank* ou *Europresse* pour la presse). Les corpus n'ont pas encore permis de systématiquement dépasser la visée normative et de se passer de la notion de langue standard. En effet, cela nécessite d'élaborer une théorie du fonctionnement des langues intégrant la variation, alors qu'une partie de la linguistique repose sur la conception plus rassurante qu'il existe une certaine homogénéité.

4.5. Édition

Il faut distinguer ici entre les interventions qui ajoutent des informations et celles qui modifient le texte lui-même. Pour les premières, on pense aux annotations dans une transcription qui précisent la durée des pauses, donnent des indications sur la prononciation d'un mot et à tous les ajouts qui vont permettre un traitement plus performant (du moins sur le plan informatique) des données (voir ci-après la section 5.3.2. sur l'annotation). On s'interrogera ici sur les interventions du second type, celles qui modifient le texte lui-même.

4.5.1. La place des variantes graphiques

Pour les corpus écrits, la question des variantes graphiques est un aspect bien connu des philologues et plus généralement des historiens de la langue. Mais depuis que l'orthographe française est devenue une institution et la même pour tous, il est difficile de ne pas s'y conformer de manière scrupuleuse. À cela s'est ajoutée la rigidité des outils

informatiques qui ont du mal à gérer la variabilité des graphies. Il existe bien quelques logiciels permettant de travailler à partir de graphies non normalisées, à l'image du lemmatiseur *LGeRM* qui a été conçu pour gérer les variations graphiques historiques du français. Mais avec la majorité des outils d'annotation automatique, l'absence de normalisation graphique dégrade fortement la qualité.
Nous faisons donc face à une situation paradoxale. D'un côté, il est indispensable de représenter aussi fidèlement que possible les données mises à disposition dans un corpus et ainsi de ne pas intervenir sur la source pour la modifier. D'un autre côté, l'absence de normalisation des graphies rend l'exploitation complexe. La solution retenue est généralement de recourir à deux versions du même corpus : une première version *in extenso* fidèle aux graphies d'origine (version diplomatique) et une seconde version en orthographe normalisée pour faciliter les recherches. Cette technique est généralement utilisée pour informatiser les copies d'élèves, mais rarement pour les textes issus du web. Cela oblige donc, dans ce second cas, à chercher l'ensemble des variantes lorsqu'on formule une requête.
Dans le cadre des corpus oraux, l'obligation d'arrêter des conventions de transcription pousse à se positionner sur les graphies à adopter et sur le lien entre la prononciation et la graphie. Cette question est encore plus brulante dans les enregistrements de jeunes enfants où la « bonne » transcription n'est pas toujours décidable. Dans certains corpus, on ne tolère presque aucune variante basée sur la prononciation ou bien on recourt à une couche phonétique pour en rendre compte. On peut également utiliser le principe des multi-transcriptions quand le choix entre au moins deux variantes est indécidable (ex : *j'étais / j'ai été*). Dans d'autres corpus, la prononciation est prise en compte et la graphie peut être différente. Ainsi dans *CLAPI,* on peut trouver la forme « dessous » écrite tantôt *d`ssous* tantôt *dessous*, la forme « autre » tantôt *aut`* tantôt *autre*, etc. De même, pour « tu », transcrit *tu* ou *t'* en fonction de la prononciation.

4.5.2. Interventions sur les textes originaux

Il est sans doute utile de distinguer entre les corpus écrits qui rassemblent des documents préexistants et les corpus oraux qui changent de médium et passent d'un enregistrement à une transcription. L'important est que d'une part, les interventions respectent des conventions ou un protocole précis pour qu'elles soient stables sur l'ensemble du corpus et d'autre part, qu'elles ne soient pas rendues invisibles ou du moins que l'on conserve la trace des modifications effectuées : on préserve ainsi la qualité des données collectées et il reste toujours possible de distinguer ce que le chercheur a modifié par rapport à la source initiale.

À l'écrit :
a) Fairon et Paumier (2007) rappellent les impératifs auxquels est confronté le chercheur qui travaille sur des écrits non-standard. Leur analyse porte sur un corpus de SMS : « Pour pouvoir être utilisé, un corpus de textes doit être lisible. […] Il est donc quasi indispensable de disposer d'une transcription en forme standard pour pouvoir réellement exploiter de telles données. »
b) Pour des productions d'élèves, on peut envisager de proposer une version qui rétablit l'orthographe standard si cela est compatible avec la recherche conduite. Dans ce cas, il

est important de mettre à disposition une version image ou diplomatique qui donne accès au texte original. Dans le projet *Ecriscol*, qui rassemble des productions d'élèves du CE1 à l'université, trois versions sont disponibles : l'une normée interrogeable par concordancier, la transcription diplomatique et la version originale sous forme d'image. Le corpus *Scoledit* comporte également ces trois versions.

À l'oral :
Tous les sites de corpus oraux donnent accès aux transcriptions et plusieurs d'entre eux permettent aussi d'entendre la version sonore ou de visionner les vidéos correspondantes. Cependant, la version écrite d'un enregistrement résulte de choix liés à l'exploitation qui est prévue (Ochs, 1979 ; Blanche-Benveniste et Jeanjean, 1987, Blanche-Benveniste, 2002). Les différences entre les divers grands projets de corpus oraux portent :
– sur la graphie et le choix d'adopter (ou pas) des formes graphiques non-conventionnelles (par exemple, faut-il noter *il y a* ou *y a* (Blanche-Benveniste, 2010a), choisir *puis* ou *pis* (Dostie, 2014) ?) ;
– sur les formes oralement ambigües du point de vue grammatical (ex : l'homonymie entre le futur et le conditionnel à la 1ère personne du singulier [>Notice futur]), où souvent le transcripteur fait un choix sans même s'en rendre compte ;
– sur la fréquence des erreurs ou des corrections involontaires (dans *OFROM*, certains passés surcomposés du type *j'ai eu fait* ont été transcrits comme des passés composés) ;
– sur les marqueurs de segmentation (doit-on reprendre des signes comme le point ou la virgule déjà installés pour l'écrit ou faire appel à d'autres signes comme # ? (Cappeau et Gadet, 2021)).
Tous ces choix ont une incidence forte car les analyses se font principalement à partir de la transcription. Ainsi, un *qu'il* transcrit *qui* (en lien avec la prononciation) ou un *t'as vu* dans lequel le *t'* est analysé comme un pronom complément risquent fort de mener à des analyses sans fondement.

4.5.3. Reformulation des textes

Deux cas doivent être bien distingués. Lorsque la reformulation n'est pas le fait du chercheur mais est une pratique sociale qui existe en dehors de la recherche elle-même, le corpus composé de ces textes reformulés doit contenir cette information dans ses métadonnées. C'est une pratique courante pour l'élaboration de supports transposés en FLE (à partir de sources authentiques) ou de textes littéraires simplifiés ou réécrits pour un public d'enfants. L'important est que l'existence de ces interventions soit dument signalée pour que les utilisateurs du corpus ne soient pas induits en erreur.
La question est différente lorsque la reformulation est le fait du chercheur lui-même qui, pour des raisons parfois dictées par des nécessités juridiques, choisit de transformer un texte original (par exemple pour mieux assurer l'anonymisation des locuteurs). Habituellement, ce souci passe par le masquage de certains noms propres (noms de personnes ou de lieux) ou par la substitution (ainsi Patricia peut devenir Laetitia, etc.). Les modifications peuvent toutefois aller au-delà comme indiqué dans Auriac-Slusarczyk et Delsart (2021) :

> « En revanche, les propos originaux et singuliers issus de la confrontation croisée entre médecins n'ont pu et ne peuvent être publiés ; les médecins ont confié verbalement

> leurs avis authentiques, au prix d'être assurés qu'on ne divulguerait cette vérité du verbe à quiconque. Ils n'ont donc pas consenti à la divulgation en l'état de leurs verbalisations. Nous avons dû gloser leurs propos, mais ne pouvons/pourrons en faire une étude trop fouillée. »

Dans le cas d'échanges confidentiels, le chercheur glose une production à laquelle il ne peut donner accès. Le corpus est-il alors constitué des propos des médecins ou des reformulations élaborées par le chercheur ? Se pose la question de la validité et du contrôle des reformulations effectuées et finalement du matériau linguistique qui est proposé au lecteur : peut-on considérer que l'analyse porte sur les propos des médecins ?

4.6. Formats

Les corpus au format texte brut, comme leur nom l'indique, ne contiennent que des chaines de caractères sans mise en forme (gras, italique) et sans information additionnelle. Juste le texte et rien que le texte. Ce format peut s'avérer inadapté pour mener à bien certaines recherches car l'intégralité des données se trouve sur le même plan. Cela signifie que pour une portion de texte donnée, les outils de recherche informatisés ne peuvent pas déterminer s'il s'agit d'un titre, d'un commentaire ou d'une note de bas de page.
L'autre format, sans doute le plus fréquemment utilisé à l'heure actuelle, est le *XML* (*eXtensible Markup Language*). Le *XML* est un langage de balisage, comme le *HTML* utilisé pour l'élaboration de sites web, qui permet de délimiter des chaines de caractères dans un document textuel. Concrètement, des zones de texte sont entourées de balises pour marquer explicitement leur contenu. Ci-dessous, un extrait du corpus *Mariage pour Tous* (*MPT*) élaboré par Nicolas Legrand. Ce corpus est composé des séances de débats qui ont eu lieu au sein de l'Assemblée nationale française avant le vote pour l'ouverture du mariage à l'ensemble des personnes quel que soit leur sexe.

```
<sp who="Hervé_Mariton">
  <speaker>Hervé Mariton</speaker>

<mpt:metadata auteur="Hervé_Mariton" interventiontype="intervention"
politicalgroup="UMP"  vote="contre" gender="male" politicalgender="UMP_male"
wing="right" winggender="right_male" debat="mpt" subject="mpt">
  <p> Tous les ministres, manifestement, n'ont pas la même position.</p>
<p>Mesdames les ministres, j'aimerais au moins que vous nous précisiez
quelle est la part au sein du Gouvernement de ceux qui sont pour la GPA et
de ceux qui sont contre, outre les trois ministres que j'ai cités. Mais
vraiment, monsieur le président, vous pourriez convier le Premier ministre
pour qu'il vienne préciser la position du Gouvernement. <mpt:interruption
type="Applaudissements" text="Applaudissements sur les bancs du groupe UMP."
/></p>

</mpt:metadata>

</sp>
```

Figure 1. Extrait en *XML* du corpus *MPT*

Ces balises permettent de savoir qui prend la parole, son groupe politique, son genre, son vote et on dispose également d'informations paralinguistiques comme les applaudissements. Le grand avantage de ce marquage explicite, c'est que cela permet de

formuler des requêtes à partir des informations contenues dans les balises et de distinguer le contenu langagier des informations paralinguistiques et documentaires.
Pour des besoins d'interopérabilité et pour ne pas systématiquement « réinventer la roue », il existe des initiatives de mise à disposition de « grammaires » de documents, à savoir un langage contrôlé destiné à lister l'ensemble des balises à utiliser, leur agencement ainsi que leur signification. Concrètement, il s'agit d'une liste de noms de balises accompagnée de leurs valeurs possibles et de leurs contraintes. Cela permet ainsi de délimiter les phrases, les paragraphes ou les tours de parole à l'aide de balises univoques et homogènes, de mentionner la date de naissance d'un auteur, etc. C'est le cas par exemple de la *Text Encoding Initiative* (*TEI*). Ainsi, si l'on souhaite informatiser une pièce de théâtre ou un roman, il suffit de se rendre sur le site de la *TEI*, de sélectionner la grammaire de document qui correspond à ses besoins et de l'appliquer.
Du côté des corpus oraux, plusieurs logiciels d'assistance à la transcription génèrent du *XML* en sortie. D'autres produisent des formats spécifiques. C'est le cas du format *TEXTGRID* utilisé par le logiciel *PRAAT* qui correspond à un format texte structuré. Le format *CHAT* du logiciel *CLAN* est également spécifique.
Du côté des métadonnées, la question de l'interopérabilité est aussi au cœur des problématiques contemporaines. Des initiatives sont encore en cours en 2025 pour définir un jeu minimal de métadonnées pour les corpus oraux comme écrits. Les métadonnées sont généralement en *XML* sous la forme d'un entête *TEI* directement au début du fichier ou bien dans un fichier tabulaire externe contenant une métadonnée par colonne.

5. TRAVAILLER AVEC DES CORPUS

Comme dit auparavant, travailler avec des corpus suppose une position scientifique claire vis-à-vis des données : souci de disposer de données *attestées*, conviction que les faits linguistiques authentiques présentent un intérêt pour les analyses, orientation du travail vers les faits plus que les exemples (l'opposition *datum* vs *exemplum* de Laks signalée en 1.1.). Définir ce qu'est un « fait » linguistique est d'ailleurs une question à la fois centrale et non consensuelle. Pour les linguistes travaillant à partir de corpus, un fait émane forcément de données attestées. L'utilisation de corpus ne suscite plus, de nos jours, les réserves d'une partie de la communauté, on peut même avoir l'impression d'une approche consensuelle, mais en réalité les pratiques ne sont pas uniformes (voir 2.4.).

5.1. Où trouver des corpus ?

Les corpus oraux existants sont désormais en grand nombre, mais certains ne sont pas accessibles pour la communauté scientifique. Cela est dommageable pour les jeunes chercheurs qui parfois sont obligés de collecter de nouveau des données similaires. Il y a des raisons à cela : transcrire est très chronophage et n'est pas toujours considéré comme une activité scientifique à part entière ; certaines situations (le milieu médical par exemple) ne permettent pas la diffusion des données. Toutefois, de plus en plus de corpus sont mis à disposition directement sur Internet.
Certains corpus à « spectre large », même s'ils ont été conçus pour une recherche précise, ont l'ambition de pouvoir être partagés et exploités dans de multiples disciplines. On

pense notamment aux corpus issus de grands projets qui rassemblent une masse importante de données.
Plusieurs corpus sont accessibles à travers des sites web qui permettent de formuler des requêtes (avec ou sans inscription préalable). D'autres peuvent être téléchargés et interrogés localement sur un ordinateur par l'intermédiaire d'un outil dédié. Il existe plusieurs sites internet recensant les corpus disponibles en langue française. On peut citer :
le site personnel de Denis Apothéloz (https://perso.atilf.fr/apotheloz/corpus/),
l'*EGF* (http://encyclogram.fr/util/liens.php),
Corpus finder (https://www.corpusfinder.ugent.be/corpora).
Un inventaire assez complet, pour les corpus écrits, est également disponible à l'adresse https://corli.huma-num.fr/inventaire-des-corpus-ecrits/ et à l'adresse suivante pour les corpus oraux : http://ircom.huma-num.fr/site/corpus.php. Il existe aussi des entrepôts à corpus comme *Ortolang* (https://www.ortolang.fr), *Nakala* (https://nakala.fr) et *CoCoON* (https://cocoon.huma-num.fr).
Les formats et les choix de conventions de transcription peuvent empêcher d'exploiter ensemble plusieurs corpus directement. C'est la raison pour laquelle des initiatives ont vu le jour pour mutualiser des corpus existants. C'est le cas du projet *Orféo* qui a abouti au *CEFC* et qui a consisté en partie à rendre compatibles des données hétérogènes. Cela a nécessité différentes étapes de correction, dans le cas des transcriptions d'oral, de normalisation des formats et d'adoption de conventions de transcription identiques. Ce projet a permis des réflexions sur la définition d'un format commun pour les transcriptions de français parlé et pour les métadonnées, et de principes de transcriptions unifiés, avec une préoccupation d'ordre qualitatif.
Dans cette même perspective de mutualisation, il existe le projet *E-Calm* (http://e-calm.huma-num.fr/) regroupant un grand nombre de copies d'élèves préalablement disséminées sur différentes plateformes. Ce dernier projet met à disposition les transcriptions *in extenso* des copies couplées à une version en orthographe standard ainsi que l'ensemble des procédés d'éditions des jeunes scripteurs (ratures, insertions, etc.), voire les annotations des enseignants.

5.2. Le web comme corpus

Avec l'arrivée d'internet, une nouvelle source de données linguistiques nativement informatisée et facile d'accès a vu le jour. Il est donc compréhensible que les linguistes s'y soient intéressés très tôt, d'autant que sa taille a crû de manière exponentielle. La question de savoir si des données extraites du web pouvaient être considérées comme un corpus a été discutée à plusieurs reprises (Emirkanian et Fouqueré 2003, Kilgarriff et Grefenstette 2003, Leech 2015, Mayaffre 2010a). Comme souvent, il est indispensable d'en mesurer les avantages (accès à une masse de données gigantesque) et les inconvénients (sources peu contrôlables : est-ce un locuteur natif ? est-ce une parodie ? une citation ? une traduction ? certains passages ont-ils été dupliqués, etc.).
En dehors d'une liste de sites bien délimitée (presse, institutions), il est impossible de maitriser parfaitement la provenance et la composition de corpus issus du web. La principale information fiable dont on dispose est représentée par le domaine (.fr ; .ch ; .be ; .qc). On pourrait penser que la taille permet de neutraliser le problème de l'hétérogénéité des données et de l'absence de traçabilité à partir du raisonnement

suivant : plus le volume est important, plus le bruit est négligeable. Concernant la qualité et le contenu, une étude sur des corpus multilingues a été menée (Kreutzer *et al.*, 2022). Celle-ci invite clairement à la prudence en mettant en avant que certains corpus web comportent moins de 50 % de phrases de qualité acceptable.

Toutefois, il faut bien reconnaitre qu'au début des années 2000 la constitution de corpus à partir du web a rendu bien des services. Elle a notamment permis la réalisation du *Corpus Évolutif de Référence du Français* (*CERF*) sous la direction de Jean Véronis, alliant le choix des sources, l'équilibre, la diversité et la disponibilité des données. Ce corpus était précurseur pour le domaine français. À une époque où les données disponibles étaient en nombre réduit, ce corpus de 10 millions de mots comportant 10 tranches homogènes, dont une de français parlé (issue de *Corpaix*, corpus constitué à Aix-en-Provence), a permis de mener à bien des recherches impossibles à faire autrement (sur la désambigüisation lexicale automatisée, par exemple). Malheureusement, le *CERF* n'a jamais pu être diffusé à cause de questions juridiques non-résolues.

D'un autre côté, on assiste ces dernières années à une montée en puissance de l'exploitation des données issues de réseaux sociaux. *Twitter* était l'une des sources les plus utilisées jusqu'à ce que la collecte de tweets soit bloquée. Mais comme pour le web en général, il était difficile d'avoir des informations précises sur les personnes à l'origine des tweets. Il convient donc d'utiliser avec prudence de telles données et d'être conscient de leurs limites. De plus, l'accessibilité des données sur les réseaux sociaux n'est pas pérenne et il est courant de voir se refermer la possibilité de les récupérer (comme dans le cas de *Twitter*).

Le *CERF* montre clairement les bénéfices que l'on peut tirer d'internet à partir du moment où le concepteur du corpus opère une sélection précise et réfléchie des données. Pour l'anglais, on trouve notamment les exemples du *Corpus of Contemporary American English* (*COCA*) et du *Corpus of Historical American English* (*COHA*) élaborés par Mark Davies. Le premier fait un milliard de mots et le second près de 500 millions. On aimerait disposer de tels corpus diversifiés pour le français. Il existe tout de même le *FrWaC* (1,6 milliard de mots) et l'*Araeneum Francogallicum III* (10 milliards de mots). Mais ces derniers sont de vastes corpus sans échantillonnage précis.

5.3. Les logiciels : puissance et limites

Depuis les années 1990, la linguistique de corpus a connu une explosion du nombre de logiciels disponibles pour la transcription des données et leur exploitation, en parallèle de l'explosion du nombre de corpus. Ci-dessous, nous faisons un tour d'horizon non-exhaustif des logiciels existants en soulignant leurs avantages et leurs inconvénients.

5.3.1. Les logiciels pour l'établissement des données

Il existe actuellement de nombreux outils informatiques destinés à la transposition automatique d'un format papier à un format numérique, et à l'accompagnement du travail de transcription pour le français parlé. Pour les textes écrits plus ou moins anciens, quand cela est possible, ils sont scannés et on leur applique un outil de reconnaissance optique de caractères (ou *OCR*). C'est ainsi qu'ont été obtenus les ouvrages disponibles sur *Google Livres* ou sur le site *Gallica.fr* de la *Bibliothèque nationale de France*. Pour

Frantext, le travail a commencé à partir de cartes perforées puis par saisie manuelle des textes sur ordinateur. Il a été adapté au gré des évolutions technologiques et se fait aujourd'hui par *OCR*, vérification humaine et balisage des textes avant mise en ligne.
Pour la presse, des sites comme *Retronews* ont eu recours aux techniques décrites ci-dessus, mais il arrive de plus en plus, pour la presse contemporaine, que les données soient récupérées nativement sous format numérique, comme sur le site *Europresse*. Il existe également des outils permettant d'aspirer directement à partir du web de grandes quantités de données pour faire des corpus sur des sujets ciblés (à partir de mots-clés) ou sur un nombre limité de sites. On appelle cela le *web scraping*. Des corpus comme *FrWaC* ont été obtenus de cette manière. De même, des logiciels enregistrant les saisies au clavier ou l'écriture manuscrite sur des tablettes ont vu le jour. Grâce à eux, il est désormais possible d'observer la production textuelle en temps réel (Cislaru et Olive, 2018).
Pour les corpus oraux, dans les années 1970, la transcription manuscrite d'un enregistrement était la norme, généralement à la machine à écrire, ce qui nécessitait un temps considérable. Parfois, on utilisait une pédale qui permettait de lancer ou d'arrêter l'écoute. Depuis les années 1990, des logiciels d'assistance à la transcription ont vu le jour et ont permis de disposer de corpus oraux informatisés et alignés texte/parole. La plupart d'entre eux sont basés sur un système multicouche de transcriptions, chaque ligne (appelée *tier*) représentant un niveau de transcription (orthographique, phonétique, parties du discours, etc.).
Les logiciels les plus utilisés à l'heure actuelle pour procéder aux transcriptions sont les suivants : *Transcriber, Praat, ELAN, CLAN*. *Transcriber* est un logiciel permettant simplement de transcrire et d'aligner manuellement la transcription sur le signal sonore. *Praat*, quant à lui, est un outil utilisé principalement par les chercheurs s'intéressant à la phonétique et à la prosodie car il permet de visualiser le signal et le tracé de diverses courbes prosodiques. Il permet également une transcription alignée au phonème ou à la syllabe grâce à des scripts comme *EasyAlign* (ce qui nécessite tout de même une correction manuelle minutieuse). *ELAN* est proche de *Praat*, mais il est plutôt utilisé pour transcrire des vidéos dans le cadre d'approches interactionnelles et conversationnelles multimodales. *CLAN* est principalement utilisé par les chercheurs s'intéressant à l'acquisition du langage.
On le voit, le choix d'un logiciel particulier dépend des objectifs de recherche et du domaine dans lequel on se situe. Plus récemment, *Transcriber JS* s'est inspiré de quelques-uns des outils cités ci-dessus et représente une alternative à *Transcriber*, qui ne permet pas de transcrire des vidéos et n'est plus mis à jour.
Chaque logiciel produit des transcriptions ayant un format spécifique. Heureusement, il existe des outils de conversion automatique permettant de passer facilement d'un format à un autre, comme *TeiConvert* (https://ct3.ortolang.fr/teiconvert/index-fr.html). Cela ouvre la possibilité d'utiliser plusieurs logiciels de transcription à la suite, en tirant bénéfice des spécificités de chacun d'eux.
Pour des enregistrements de bonne qualité avec peu voire pas de chevauchements de parole, il est envisageable d'utiliser des logiciels de transcription automatique moyennant une vérification manuelle approfondie. Une expérience de ce type est présentée dans Tancoigne *et al.* (2020).

Du côté des métadonnées, chaque logiciel a ses propres exigences. Par exemple, dans *CLAN*, celles-ci sont prédéfinies et précédées en début de transcription par le caractère @. Au sein du consortium *Corli*, un logiciel (*TeiMeta*) a été développé pour uniformiser le format et le contenu des métadonnées. Le principe est que les projets de constitution de nouveaux corpus partent d'un jeu de métadonnées identique et puissent en ajouter d'autres s'ils le jugent nécessaire.

5.3.2. Annotation automatique

Une fois le corpus établi, se pose la question des annotations automatiques ou manuelles. De nombreux logiciels d'annotation en morphosyntaxe existent et il devient de plus en plus facile de lemmatiser et d'annoter automatiquement en parties du discours. La qualité de ces annotations est généralement assez bonne (supérieure à 90% de précision), mais cela ne dispense pas de bien connaitre le jeu d'étiquettes retenu et les critères appliqués pour leur attribution. D'autant que pour certaines parties du discours, les résultats peuvent s'avérer discutables. Il faut donc rester vigilant et ne pas utiliser sans précaution des corpus annotés. Une solution consiste à vérifier et à corriger les étiquettes à la main, quand cela s'avère possible.

L'annotation automatique possède l'inconvénient d'imposer au linguiste une analyse des données avant que le travail ait débuté. En effet, qu'il s'agisse d'une segmentation en phrases ou d'une annotation en parties du discours, il y a forcément une orientation théorique ou pratique qui peut invisibiliser ou dénaturer certains phénomènes.

Pour prendre un exemple trivial, la distinction entre nom et adjectif peut fluctuer d'un logiciel à un autre. Autre exemple : *faux* dans *il chante faux* pourra être annoté comme un adjectif ou comme un adverbe en fonction de choix théoriques. Et le logiciel d'annotation *TreeTagger* a la fâcheuse habitude de lemmatiser systématiquement *suis* par *être*|*suivre*. Sur ce dernier cas, si l'on n'y prend pas garde, lorsque l'on va rechercher toutes les occurrences renvoyées par le lemme *être*, on ne récupèrera jamais *suis*. Il y a également la sous-détermination du jeu d'étiquettes. Certains logiciels peuvent avoir tendance à étiqueter *bon* comme adjectif. Or, dans les productions orales, cette étiquette se révèlera fausse dans 90 % des cas (Valli et Véronis, 1999) étant donné que *bon* est fréquemment utilisé comme une particule (ex : *bon tu viens*). Dans le modèle *Universal Dependencies* (UD), il n'y a pas de distinction entre auxiliaire, verbe attributif et verbe plein pour *être*. Il importe a minima qu'une telle option (qui peut résulter de choix justifiés sur le plan théorique) soit connue des utilisateurs. Avant de lancer une requête, il faut donc lire en détail le manuel d'annotation associé au logiciel pour établir des correspondances entre l'analyse sous-jacente au système d'annotation et l'objectif de sa propre recherche.

Du côté des logiciels d'annotation syntaxique, la technologie s'améliore sans cesse et leur usage est sans doute amené à se généraliser. En 2025, la qualité des annotations ne nous semble pas suffisante pour envisager une exploitation directe. Toutefois, un corpus comme le *French Treebank*, qui a été vérifié manuellement, se prête à des études linguistiques riches et diversifiées (Thuilier, 2012).

Pour toutes ces raisons, les linguistes contemporains travaillent rarement à partir de corpus bruts. Les traitements les plus courants sont la lemmatisation et l'annotation en parties du discours. Les logiciels informatiques existants pour effectuer ces tâches sont

faciles d'accès et d'utilisation. Et de plus en plus de corpus disposent d'une annotation syntaxique avec les relations de dépendance ou les constituants.
En conclusion, l'annotation automatique présente un intérêt certain pour extraire des exemples plus rapidement et peut se révéler fiable dans certains cas. Mais elle ne doit pas être considérée comme une information valide à suivre aveuglément. Il faut en questionner les limites et les biais éventuels.

5.3.3. Interfaces de consultation / langages de requête

Certains corpus sont disponibles avec leur propre outil d'interrogation *: Frantext, CFPP, ESLO, CEFC, OFROM, SCIENTEXT* (liste non-exhaustive). Dans d'autres cas, majoritaires, seuls les textes sont disponibles. Si la disponibilité d'une interface de requête peut se révéler un avantage, cela peut rendre difficiles les comparaisons entre des résultats provenant de corpus distincts. Les informations extraites peuvent être de nature très différente. Par exemple, l'interface d'*ESLO* donne le nombre de segments qui contiennent au moins une occurrence du motif recherché. Il suffit qu'il y ait plusieurs motifs dans un même segment pour que le décompte ne corresponde pas au nombre d'occurrences.
L'export des résultats est également assez varié et pas toujours paramétrable. On peut donc manquer de contexte ou bien ne pas disposer des données sous la forme qui nous convient pour l'exploitation visée. Par exemple, si l'occurrence n'est pas isolée dans une colonne au format *KWIC*[2], cela va rendre impossible les tris sur les contextes gauche ou droit. De même, on ne pourra pas forcément récupérer une métadonnée pourtant indispensable à notre étude.
Pour résoudre ces problèmes, on peut télécharger les corpus et utiliser un même outil pour les interroger, comme *TXM, Sketchengine* et *CQPWeb*. Ces outils permettent d'interroger différentes couches d'annotation ainsi que les métadonnées. Dans certains logiciels (c'est le cas de *TXM*), il est possible d'écouter les fichiers audios synchronisés avec la transcription ou de lire les fichiers vidéo.
Des outils d'interrogation comme celui de *Frantext* permettent de réduire le nombre d'exemples en procédant à une sélection aléatoire. Cette fonctionnalité est très utile quand on souhaite travailler sur un nombre précis d'exemples ou quand il n'est pas envisageable d'analyser un trop grand nombre d'occurrences. Dans ce cas, la sélection aléatoire est incontournable pour ne pas biaiser les résultats en se limitant à des données qui ne sont pas représentatives de l'ensemble du corpus. Il est également possible de procéder à ce genre de sélection en utilisant les fonctionnalités proposées dans les tableurs.
Il arrive souvent qu'un logiciel dispose de son langage de requête spécifique. Cependant, le langage *CQL* (pour *Corpus Query Language*) tend à devenir un standard et permet de formuler des requêtes sur différents niveaux d'annotation. C'est ce langage qui a été retenu par *TXM* et l'interface d'interrogation de *Frantext*. Il s'agit d'exprimer ses requêtes sous la forme de couples attribut-valeur. L'attribut correspond au nom d'une propriété lexicale ou structurelle présente dans un corpus donné (la forme, le lemme, la partie du discours, l'identifiant d'un locuteur, une métadonnée) et sa valeur correspond à la forme

[2] *KWIC* signifie *KeyWord In Context*. Il s'agit du format de concordances le plus répandu dans lequel l'occurrence recherchée est isolée au centre, entourée de son contexte gauche et droit.

recherchée (*dis, dire, Verbe*). Ce langage se révèle adapté à un grand nombre de recherches en linguistique et son apprentissage est relativement rapide[3].
Il existe également des langages de requête permettant d'interroger des corpus annotés en syntaxe. Ils permettent par exemple de retrouver tous les sujets grammaticaux d'un verbe donné ou la liste des adjectifs en position d'attribut du sujet. Les plateformes *Scienquest*, *Lexicoscope* et *Grewmatch* proposent des corpus de ce type à interroger. Si *Scienquest* est très facile d'usage avec des requêtes formulées à l'aide de petites boites à relier entre elles via des relations syntaxiques, il n'en va pas de même des autres logiciels qui nécessitent généralement une formation pour en maitriser le maniement.
Après le choix d'un logiciel d'interrogation – si ce dernier n'est pas fourni avec le corpus – vient la phase d'exploitation. La situation la plus courante, pour les analyses grammaticales, est d'exporter les concordances sous forme de tableaux et de procéder à leur annotation manuelle en fonction de critères définis par le chercheur. L'étape suivante est d'utiliser les fonctionnalités de filtres des tableurs afin de dénombrer chaque catégorie ou bien d'utiliser des méthodes statistiques plus ou moins élaborées. Pour ce faire, l'environnement *R*[4] et ses librairies sont de plus en plus souvent utilisés (Desagulier, 2017).

5.4. L'unité de base d'exploitation des corpus : le mot

Le statut du mot en tant qu'unité linguistique pertinente a régulièrement été remis en cause (Pernier, 1986, Berrendonner et [Reichler]-Béguelin, 1990). Cependant, il a retrouvé une deuxième vie grâce aux corpus. Ou plutôt grâce aux logiciels informatiques qui peuvent aisément rechercher des occurrences de *mots graphiques* et plus difficilement des *morphèmes*, *lexèmes* ou autres unités qui nécessitent des traitements plus complexes. Ils peuvent tout aussi aisément les compter. On peut remarquer que même la taille des corpus oraux est le plus souvent indiquée en mots graphiques. De plus, la plupart des annotations ont pour ancrage cette unité. Qu'il s'agisse des lemmes ou des parties du discours, les étiquettes sont liées aux mots. Même si la place de cette unité est centrale pour effectuer les recherches dans les corpus, il ne faut pas perdre de vue qu'il ne s'agit pas d'une notion linguistiquement fondée. C'est une unité définie de manière pratique, particulièrement bien adaptée aux traitements informatisés.
Dans la plupart des cas, une liste de caractères séparateurs (espace, virgule, etc.) permet d'effectuer un premier découpage systématique. Et il arrive que cette segmentation soit complétée par une liste de formes à regrouper à partir d'un lexique, à l'image de *c'est-à-dire peut-être* et *quand même*. Pour éviter toute confusion, en linguistique de corpus, on préfère souvent parler de *tokens* (et non de mots), le token représentant une unité de segmentation graphique reposant généralement sur une liste de caractères séparateurs. Les caractères de ponctuation sont comptés comme des tokens. Ainsi, la taille des corpus est

---

[3] Pour une présentation de la syntaxe des requêtes *CQL*, consulter la page (et les suivantes de la section 12) : https://txm.gitpages.huma-um.fr/textometrie/files/documentation/manual/0.7.9/fr/manual60.xhtml

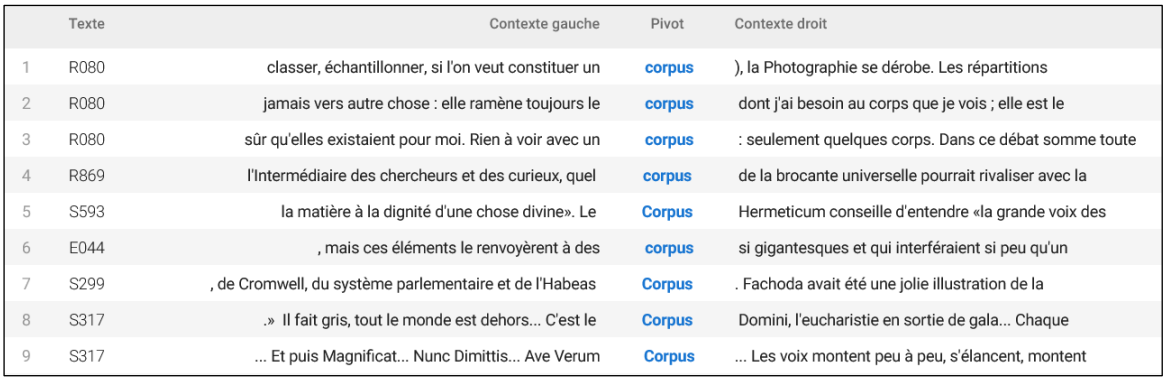
[4] *R* est un langage de programmation et un logiciel couramment utilisé pour faire des statistiques.

parfois exprimée en tokens, parfois en mots graphiques sur la base d'un découpage automatique.
Toutefois, il est important de préciser que « si « l'entrée » dans les corpus se fait par les mots, les phénomènes étudiés vont très souvent au-delà des mots et concernent des aspects syntaxiques, discursifs, sémantiques… » (Condamines et Dehaut, 2011 : 270). On a déjà signalé l'importance des cooccurrences et des collocations dont les corpus ont favorisé l'étude (Oakey, 2009). À partir des mots, il y a donc la possibilité (variable d'un corpus à l'autre en fonction du logiciel d'interrogation) de faire des recherches plus complexes : par exemple de rechercher des constructions en *j'ai X qu- Y* ou l'ensemble des formes passives. Mais cela se fait forcément avec une plus ou moins grande quantité d'erreurs et nécessite par conséquent une phase de tri manuel assez chronophage.
Un logiciel tel que *SketchEngine* (https://www.sketchengine.eu) possède un outil nommé *word sketch*. Ce dernier présente les collocations de manière compacte et facile à appréhender en tenant compte des relations syntaxiques. Dans le tableau ci-dessous, on peut voir en un coup d'œil tous les verbes dont le lexème *projet* est sujet, complément ou entretient avec celui-ci une autre relation syntaxique, et ce par ordre décroissant d'importance :

Figure 2. Capture d'écran par word sketch du token *projet* dans le corpus *Europarl spoken parallel*

5.5. La question du contexte

En syntaxe, en lexicologie/lexicographie et en sémantique en particulier, le concordancier est utilisé pour extraire une forme ou un patron et disposer rapidement des environnements textuels dans lesquels cette forme se rencontre dans le corpus. En effet, le sens des items lexicaux « dépend largement de leurs collocats, des unités pertinentes, grammaticales ou lexicales, qui les entourent » (Teubert, 2009). Les réponses aux requêtes en ligne sur un gros corpus (*Frantext* pour l'écrit ou les divers projets de corpus oraux) prennent appui sur ce même outil.
Le travail sur corpus engage à développer une linguistique qui repose sur des lignes discontinues, à savoir la fenêtre de résultats du concordancier.



	Texte	Contexte gauche	Pivot	Contexte droit
1	R080	classer, échantillonner, si l'on veut constituer un	corpus	), la Photographie se dérobe. Les répartitions
2	R080	jamais vers autre chose : elle ramène toujours le	corpus	dont j'ai besoin au corps que je vois ; elle est le
3	R080	sûr qu'elles existaient pour moi. Rien à voir avec un	corpus	: seulement quelques corps. Dans ce débat somme toute
4	R869	l'Intermédiaire des chercheurs et des curieux, quel	corpus	de la brocante universelle pourrait rivaliser avec la
5	S593	la matière à la dignité d'une chose divine». Le	Corpus	Hermeticum conseille d'entendre «la grande voix des
6	E044	, mais ces éléments le renvoyèrent à des	corpus	si gigantesques et qui interféraient si peu qu'un
7	S299	, de Cromwell, du système parlementaire et de l'Habeas	Corpus	. Fachoda avait été une jolie illustration de la
8	S317	.» Il fait gris, tout le monde est dehors... C'est le	Corpus	Domini, l'eucharistie en sortie de gala... Chaque
9	S317	... Et puis Magnificat... Nunc Dimittis... Ave Verum	Corpus	... Les voix montent peu à peu, s'élancent, montent

Figure 3. Capture d'écran de la concordance du mot *corpus* dans Frantext

On peut avoir besoin de contextes larges quand on s'intéresse aux connecteurs corrélatifs, le second élément de la corrélation pouvant se trouver très loin du premier, obligeant à aller lire les textes de manière suivie. En théorie, le contexte peut être étendu sans trop de limites, sauf dans des cas comme *Frantext* pour des questions de droits. Cependant, dès que le contexte dépasse la quarantaine de mots, les concordances deviennent difficiles à lire et à exploiter. S'il n'y prend garde, le chercheur se trouve face à un paradoxe : il dispose de ressources de plus en plus volumineuses (qu'il connait de façon de plus en plus lointaine et approximative) et finit par travailler sur des suites de lignes (soit des fenêtres très réduites) :

> « dès que le nombre des contextes est un peu élevé, les mises en contextes ainsi réalisées (comme les concordances, etc.) deviennent des objets difficilement manipulables, même sous forme informatisée. L'organisation de ces listes (définition et ordre de présentation des contextes) influence très fortement la perception de divers phénomènes relatifs à la forme-pôle ». (Habert *et al.*, 1997 : 183)

La recherche du pronom personnel *je* dans un corpus oral peut servir à illustrer la nécessité de disposer d'un contexte large pour conduire une analyse. Le corpus *DECLICS2016* (Blasco, 2022) comporte des échanges entre des patients et des psychanalystes. Dans la parole du psychanalyste, il faut constamment vérifier si le *je* est celui du professionnel de santé ou s'il s'agit en fait de celui du patient dont le psychanalyste cite les propos.
Dans les situations où on se retrouve submergé par un déluge de données, il peut être tentant de ne plus analyser directement les exemples et de produire une analyse purement statistique. Cela ne va pas sans poser certaines questions méthodologiques, que nous abordons plus loin (5.4.3.).

5.6. Comment le corpus oriente le travail

Il n'est pas envisageable de faire ici un point détaillé sur la façon dont les corpus ont pu transformer le travail en linguistique. On se contentera de quelques remarques en lien avec le choix des faits de langue observés et l'aspect quantitatif.

5.6.1. Choix des observables

Comme on l'a vu, la taille des corpus contemporains rend difficile une connaissance détaillée de leur contenu. Le recours à un concordancier permet de faire apparaitre dans leurs environnements des formes ou des patrons spécifiques. Ce mode d'obtention des données facilite grandement l'analyse systématique des distributions de nombreux lexèmes[5], mais ce type d'accès peut faire perdre des informations sur le contexte et la dimension « textuelle » du corpus. De même, les résultats peuvent parfois être difficiles à interpréter à cause du manque d'information accompagnant le corpus, ce qui peut donner lieu à des erreurs d'analyse.

La recherche de formes dans les corpus ne peut s'envisager, par définition, qu'à partir d'unités textuelles circonscrites relativement faciles à retrouver. Dès lors, des relations non marquées, comme dans l'exemple *je me suis marié j'avais 20 ans*, ne peuvent pas être extraites ni observées à l'aide d'un concordancier. De même, certaines relations anaphoriques avec des distances importantes entre les unités coréférentes sont peut-être plus délicates à observer (même si sur ce dernier point le corpus *Democrat* (Landragin, 2021) propose des solutions). Il existe ainsi un risque non négligeable de voir disparaitre du champ de la recherche certains phénomènes linguistiques impossibles à extraire automatiquement à l'aide d'outils informatiques.

Travailler à partir d'un concordancier suppose aussi que l'on ait déjà une idée assez précise de ce que l'on cherche, contrairement à la lecture/écoute de corpus qui permet la découverte de phénomènes nouveaux ou du moins que le chercheur (voire la communauté linguistique) n'avait pas encore remarqués. La lecture suivie des transcriptions ou l'écoute des enregistrements réserve donc encore de nombreuses surprises.

5.6.2. Portée des résultats

La description à partir d'un corpus peut donner lieu à des raccourcis qu'il est bon d'interroger : est-il raisonnable de prétendre décrire la langue écrite en s'appuyant uniquement sur un corpus de textes littéraires ? Décrit-on le français parlé en s'appuyant sur un seul corpus oral, même quand celui-ci présente une diversification des types de discours ? Littéralement non, même si cela ne condamne pas le travail de description entrepris sur ces bases.

> « En termes opérationnels, le terme « représentatif » signifie que l'étude d'un corpus (ou d'une combinaison de corpus) peut remplacer l'étude d'une langue entière ou d'une variété de langue. Cela signifie que toute personne menant une étude raisonnée sur un corpus représentatif (considéré comme un échantillon d'une population plus large, son univers textuel) peut extrapoler à partir du corpus à l'univers entier de l'utilisation de la langue dont le corpus est un échantillon représentatif. Mais dans l'état actuel des choses, pouvons-nous même prétendre à une « validité apparente » (pour utiliser un terme employé pour des tests linguistiques) pour la représentativité des corpus avec lesquels nous travaillons ? » (Leech, 2015 ; trad. Cappeau)

---

[5] Le site de Patrick Dendale
https://forms.uantwerpen.be/en/projects/lexicales/recherche-dans-la-bibliographie/
qui recense les publications sur des formes particulières, donne une idée de la diversité des recherches.

Divers garde-fous (sous forme de questions que se pose le chercheur) peuvent être utilisés. Par exemple, on peut se focaliser sur la taille (voir 4.1.) ou la diversité (voir 4.4.). Cela revient à se demander si le corpus possède une taille ou un contenu en adéquation avec les objectifs poursuivis. Là encore, c'est l'objectif du travail qui peut guider la collecte des exemples : les faits de langue recueillis sont-ils volontairement uniformes ou homogènes ? Le chercheur a-t-il veillé à collecter des productions diversifiées, si cela lui est nécessaire ? Habert (2000) fait un tour de la question à travers divers facteurs importants en lien avec la représentativité dans les corpus. Une mesure de prudence est de ne pas s'enfermer dans un seul corpus et d'en consulter plusieurs par principe.
Le chercheur doit avoir une vision claire et informée des caractéristiques de son corpus et de la portée des analyses qu'il va conduire. Si l'on s'intéresse aux mots en verlan en français contemporain, *Frantext* n'est sans doute pas la meilleure source pour collecter des exemples. Le corpus des formes obtenues ne sera, au mieux, représentatif que du français écrit, à majorité littéraire. Limite dont il faut être conscient. Des corpus oraux, tels que *MPF*, et écrits issus du web (avec les précautions signalées en 4.7.). peuvent aider à établir un inventaire plus en lien avec l'époque contemporaine.
En conclusion, tout corpus étant par nature limité, il faut se prémunir de la tentation d'en tirer des enseignements de portée générale en l'absence d'une réflexion poussée.

5.6.3. Approches qualitative et quantitative

En sciences humaines, il est courant de distinguer approche quantitative et approche qualitative. L'approche quantitative donne lieu à un traitement statistique plus ou moins sophistiqué : fréquence du motif recherché, évolution de cette fréquence au cours du temps, mesure de l'influence d'un (ou plusieurs) paramètre(s), etc. De son côté, l'approche qualitative accorde un soin tout particulier à la constitution des données et à une analyse descriptive fine se résumant parfois à un texte ou à un extrait de texte. Certains privilégient le qualitatif (tel que défini dans la citation ci-dessous), d'autres le quantitatif.

> « La recherche est dite « qualitative » principalement dans deux sens : d'abord, dans le sens que les instruments et méthodes utilisés sont conçus, d'une part, pour recueillir des données qualitatives (témoignages, notes de terrain, images vidéo, etc.), d'autre part, pour analyser ces données de manière qualitative (c'est-à-dire en extraire le sens plutôt que les transformer en pourcentages ou en statistiques) ; la recherche est aussi dite qualitative dans un deuxième sens, qui signifie que l'ensemble du processus est mené d'une manière « naturelle », sans appareils sophistiqués ou mises en situation artificielles, selon une logique proche des personnes, de leurs actions et de leurs témoignages (une logique de la proximité : cf. Paillé, 2007). Ainsi en est-il de l'analyse des données qui met à profit les capacités naturelles de l'esprit du chercheur et vise la compréhension et l'interprétation des pratiques et des expériences plutôt que la mesure de variables à l'aide de procédés mathématiques. » (Paillé et Mucchielli, 2012 : 13)

Plutôt que d'opposer qualitatif et quantitatif, il peut être pertinent de jouer sur leur complémentarité. Voici une illustration : l'approche qualitative favorise l'observation fine d'une portion de corpus pour dégager des pistes et des hypothèses de travail tandis que l'approche quantitative (massive) vient en renfort pour valider ou non les hypothèses en question sur des données bien plus larges. On évite ainsi le biais qu'un fragment de

corpus pourrait introduire en se donnant l'opportunité de vérifier si une généralisation est envisageable.

> « dans les approches qualitatives, l'analyse et l'interprétation se font « à l'œil » ou « à la main » ou même, pourrait-on dire, « à l'humain », alors que dans l'approche quantitative, l'analyse et l'interprétation se fait « à l'instrument » » (Paveau, 2014 : 3-4).

Signalons que certaines disciplines comme l'analyse conversationnelle revendiquent d'être moins touchées par les risques pointés ici. Dans Mondada (2017), des séquences précisément identifiées (la requête et sa satisfaction) font l'objet d'analyses « dans une approche où l'attention aux détails multimodaux refuse les généralisations précoces et privilégie plutôt la spécificité du contexte ». Mais ces approches encourent à leur tour le reproche de ne pas chercher suffisamment les généralisations.

5.6.4. Questions autour des comptages

Deux points feront, ci-dessous, l'objet de commentaires : la pratique du comptage de certains faits langagiers et le recours aux statistiques.
On pourrait reprendre la formule de Habert (déjà citée en 4.1.) « gros c'est beau » et l'appliquer au nombre d'exemples sur lesquels travaille le chercheur. L'appui sur des corpus massifs et l'automatisation des traitements ont tendance à faire enfler la quantité des observables. S'il n'y a pas de plafond quant au nombre d'occurrences à analyser, des réflexions ont eu lieu concernant le plancher minimal à respecter. Par exemple, Blanche-Benveniste (1995) considérait qu'il fallait un minimum de 50 attestations pour les exemples oraux. Ce nombre peut sembler modeste et arbitraire, mais il garantit à la fois une certaine représentativité et la possibilité de connaitre et contrôler les exemples. Sinclair (1991) proposait de fonctionner par palier de 50 exemples à observer avant de passer aux 50 exemples suivants pour voir si de nouveaux fonctionnements apparaissent ou pas. En morphologie, il arrive qu'une seule occurrence sur internet suffise pour valider l'existence d'une forme (Huguin, 2021).
En fait, ces seuils doivent être interrogés et il est difficile de définir des limites inférieures et supérieures pour le nombre d'exemples à considérer. Cela dépend largement des phénomènes observés. Avant de fixer un nombre d'exemples, encore faut-il être au clair sur certaines questions :

• Comment appréhender les exemples peu fréquents voire absents du corpus ? Une faible fréquence peut être l'indice d'un emploi accidentel ou peu productif. Un exemple rare peut aussi être vu comme le chainon manquant pour expliquer un phénomène ou une tournure en cours d'émergence. On le voit, la perspective peut être fort différente et il n'est pas évident de trancher. Par exemple, Gillet (2024) considère que le faible nombre d'occurrences de sujet clitique postposé dans les interrogatives chez de jeunes enfants est le signe que cette tournure est peu productive. Au contraire, le Groupe de Fribourg (2024) lui prédit « un bel avenir en tant que fait de système », bien qu'elle appartienne à la langue soutenue et soit peu fréquente en français parlé. On peut trouver des réflexions sur cette problématique dans Blanche-Benveniste (1995), Cappeau (2005), Rouget (2005), Bilger et Cappeau (2021) et Béguelin *et al.* (sous presse). Pour pouvoir interpréter les

signaux faibles dans un corpus, il est recommandé de contraster des données de natures différentes.

• Que faut-il entendre par 50 ou 100 occurrences ? En syntaxe ou sémantique, s'agit-il de rassembler 50 exemples d'un même lexème ou 50 exemples avec un emploi ou une valeur spécifique d'une forme ? Si l'on reprend l'exemple de *bon* (cité en 5.1.2.), recueillir à l'oral 50 occurrences de la forme *bon* ou 50 occurrences de l'adjectif *bon* ne nécessite pas le même effort. De même, que ce soit à l'écrit ou à l'oral, réunir 100 occurrences de la forme *alors que* sera plutôt rapide mais réunir 100 exemples de *alors que* à valeur temporelle deviendra nettement plus compliqué (puisque dans son usage contemporain, cet élément grammatical a, pour l'essentiel, une valeur adversative).

• L'élément comptabilisé est-il clairement défini ? Le fait de compter et de manipuler des nombres peut parfois créer une impression de rigueur scientifique dont doit se prémunir le chercheur. Celui-ci a normalement réfléchi en amont à « l'objet » qu'il veut comptabiliser. Ainsi, dénombrer les phrases d'un texte peut constituer une entrée intéressante lorsqu'elle concerne des productions normées publiées (littéraire, journalistiques…). Mais cette approche ne peut pas être, de façon automatique, transposée si elle concerne des productions non normatives (textes d'élèves, forums, etc.) dans lesquelles la ponctuation est très souvent lacunaire ou ne joue plus le rôle qu'on lui attribue habituellement à l'école. De même, une annotation automatique ou manuelle remplie d'erreurs ou basée sur des catégories mal définies va tout de même permettre d'obtenir un résultat quantitatif. Mais ce résultat sera pour le moins discutable. Bref, on peut probablement tout compter mais l'important est tout de même de savoir précisément ce que l'on compte.

Il ne faudrait toutefois pas sous-estimer ce que les comptages peuvent apporter à notre connaissance de la langue. Dans leur grammaire, Biber *et al.* (1999) indiquent comment se répartissent quantitativement les différents phénomènes qu'ils décrivent dans les corpus sur lesquels ils s'appuient. Cette orientation modifie profondément la présentation des ouvrages décrivant la langue : au lieu d'accorder une place démesurée à des faits peu fréquents dans l'usage courant, on hiérarchise davantage les phénomènes en fonction de leur importance quantitative. Le nombre vient alors tempérer l'influence de la tradition ou la subjectivité du linguiste et remet ainsi à sa juste place des exemples du type *Marie mange une pomme et Pierre une poire*, très peu fréquents dans les corpus oraux.
Un article de la revue *La Recherche* (n° 572, janvier/mars 2023, *Pourquoi les sciences sociales ont du mal avec le quantitatif*) se penchait sur l'intérêt des approches statistiques trop peu exploitées par les sciences sociales. La situation en linguistique est plus nuancée. Charles Muller (1993) a été un pionnier dans le domaine français. Et le développement des corpus s'est accompagné d'un recours plus important aux statistiques (Gries, 2021). Celles-ci occupent donc une place notable dans le champ.
Toutefois, la « magie » des nombres et le caractère objectif de la quantification des données doivent sans doute être questionnés. La sophistication des méthodes statistiques, tout comme la séduction des graphiques de synthèse ou l'appui sur des logiciels de traitement des données (voir 5.3.3.), doivent constamment rester sous le contrôle du chercheur. Les outils statistiques ne peuvent se substituer à la réflexion et aux questions

linguistiques qui sont la base de la recherche. Pour limiter le risque d'avoir des résultats statistiquement solides mais linguistiquement discutables, il faut pouvoir vérifier les données directement. Heureusement, les pratiques évoluent et il arrive de plus en plus souvent que les données annotées soient librement disponibles pour permettre de détecter les erreurs et les éventuels biais.

6. BILAN PROVISOIRE

Au terme de ce rapide parcours autour des corpus et de leur exploitation, il apparait que ce champ a désormais pris une place importante en linguistique française. De plus en plus de travaux scientifiques sont publiés en prenant pour source unique ou principale des données langagières authentiques, écrites comme orales. Ces approches, en pleine évolution, n'en demeurent pas moins assez hétérogènes. Coexistent différentes tendances allant de l'usage de statistiques extrêmement élaborées à des comptages largement « artisanaux ». De même, la nature des données, leur taille ainsi que le degré d'automatisation des traitements sont très variés, avec un retour au texte intégral non-systématique. Tout ceci interroge la validité des résultats obtenus et la place prise par la technique au détriment éventuel de la rigueur philologique. La technologie a permis des avancées prodigieuses, permettant de traiter des volumes de textes impossibles à observer « à l'œil nu ». Mais ces progrès comportent des contreparties et invitent à contrôler et maitriser, autant que possible, les différentes étapes du travail de description.
Pour finir, nous aimerions souligner deux tendances saillantes pouvant nuire à l'essor de la linguistique de corpus :

1. Un temps considérable est passé à commenter la constitution du corpus au détriment de l'analyse linguistique proprement dite. C'est une tendance semble-t-il récente qui reflète l'engouement autour des corpus mais aussi le danger : le corpus semble le point de départ et d'aboutissement du travail du chercheur. De nombreux travaux se contentent aussi de dénoncer le fait que l'on n'a pas assez de données en français. Même si l'on peut souhaiter disposer de données (toujours) plus nombreuses et variées, il est sans doute désormais possible de passer à une description d'envergure avec les données existantes.

2. Malgré des mises en garde anciennes, certaines études descriptives sur corpus se font sans forcément tenir compte des conséquences, sur les résultats, d'une mauvaise connaissance des données, des conventions de transcription ou des choix d'annotation (Béguelin *et al.*, sous presse). Il est utile de rappeler que l'établissement du corpus fait déjà partie du travail du chercheur (pour l'oral, la transcription puis la correction de ce premier état sont incontournables) et que ce dernier ne peut se dispenser de réfléchir aux différentes questions soulevées durant cette phase. Rastier rappelle les étapes du travail à respecter pour que l'utilisation des corpus soit méthodologiquement acceptable :

> « (i) analyse de la tâche et production des hypothèses ; (ii) constitution d'une archive et sélection d'un corpus de référence ; (iii) élaboration des corpus de travail ; (iv) traitement instrumenté de ces corpus, en contrastant corpus de travail et corpus de référence ; (v) interprétation des résultats et retour aux sources textuelles pour valider l'interprétation. » (Rastier, 2011 : 13).

7. BIBLIOGRAPHIE

Références importantes

Aarts, Bas. 2000. Corpus Linguistics, Chomsky, and fuzzy tree fragments. In *Corpus Linguistics and Linguistic Theory*. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora, C. Mair & M. Hundt (eds), 5-13. Amsterdam: Rodopi.

Baker, Paul (ed). 2008. *Contemporary Corpus Linguistics*. London/New-York: Continuum.

Bilger, Mireille (éd.). 2000a. *Linguistique sur corpus – Études et réflexions*. Paris/Perpignan : Presses universitaires de Perpignan.

Bilger, Mireille (éd.). 2000b. *Corpus. Méthodologie et applications linguistiques*. Honoré Champion.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan and Finegan, Edward. 1999. *Longman grammar of Spoken and Written English*. London: Pearson.

Blanche-Benveniste, Claire et Jeanjean Colette. 1987. *Le français parlé. Transcription et édition*. Paris : Didier Érudition.

Firth, John Rupert.1957. A synopsis of linguistic theory, 1930-55. *Studies in linguistic analysis*, Special Volume of the Philological Society, Oxford: Blackwell, 1-31.

Habert, Benoît, Nazarenko, Adeline et Salem, André. 1997. *Les linguistiques de corpus*. Paris : Armand Colin.

Habert, Benoît. 2005. Portrait de linguiste(s) à l'instrument. *Texto!*, vol. X, n°4. http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Habert, Benoît. 2006. *Instruments et ressources électroniques pour le Français*. Paris : Ophrys.

Laks, Bernard. 2010. La linguistique des usages : de *l'exemplum* au *datum*. *L'exemple et le corpus : quel statut ? Travaux linguistiques du CerLiCO*, 23, 11-26.

Léon, Jacqueline. 2008. Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages*. 171 : 12-33. https://www.cairn.info/revue-langages-2008-3-page-12.htm.

McEnery, Tony and Hardy, Andrew. 2011. *Corpus Linguistics*. Cambridge Textbooks in Linguistics, Cambridge University Press.

O'Keeffe, Anne and McCarthy, Michael J. 2022. *The Routledge Handbook of Corpus Linguistics*. Routledge, Second edition.

Ochs, Elinor. 1979. Transcription as Theory. In Elinor Ochs, Bambi B. Schieffelin (eds). *Developmental Pragmatics*. New York: Academic Press, 43-72.

Poudat, Céline et Landragin, Frédéric. 2017. *Explorer un corpus textuel. Méthodes - pratiques - outils*. De Boeck Supérieur.

Rastier, François. 2004. Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd). *La linguistique de corpus*. Rennes : PUR, 31-46. http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.

Sinclair, John. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2004. *Trust the Text – language, corpus and discourse*. New York. Routledge.

Teubert, Wolfgang. 2009. La linguistique de corpus : une alternative. *Semen*, 27, 185-211. (Traduction de Valérie Lebaud). https://doi.org/10.4000/semen.8914.

Viana, Vander, Zyngier, Sonia and Barnbrook, Geoff (eds). 2011. *Perspectives on Corpus Linguistics*. Volume 48. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Zufferey, Sandrine. 2020. *Introduction à la linguistique de corpus*. ISTE Éditions.

Revues consacrées aux corpus

Corpus : https://journals.openedition.org/corpus/

Corpus Linguistics and Linguistic Theory : https://www.degruyter.com/journal/key/cllt/html?srsltid=AfmBOop4maJwsyboW_tddke4PBzDnBdrxmel0c4UTSsoZ7C4fT1AUNoH

International Journal of Corpus Linguistics : https://benjamins.com/catalog/ijcl?srsltid=AfmBOop3tAsfjHXSC7D1K8nAiilQk3LeyicXeDM92baOJmTZwc6Zp5kk

Numéros de revue consacrés aux corpus

Corela, Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage, 2017 HS-21 : https://journals.openedition.org/corela/4797

Histoire Épistémologie Langage, Constitution de corpus linguistiques et pérennisation des données, 2016/2 tome 38 : https://www.persee.fr/issue/hel_0750-8069_2016_num_38_2

Langages, Construction des faits en linguistique : la place des corpus, 2008/3 n°171 : https://shs.cairn.info/revue-langages-2008-3

Langages, L'analyse de corpus face à l'hétérogénéité des données, 2012/3 n°187 : https://shs.cairn.info/revue-langages-2012-3

Langue française, Collocations, corpus, dictionnaires, 2006/2 n°150 : https://shs.cairn.info/revue-langue-francaise-2006-2

Éla. Études de linguistique appliquée, Linguistique de corpus appliquée, 2017/4 n°188 : https://shs.cairn.info/revue-ela-2017-4

Revue française de linguistique appliquée, Corpus : bilan et perspectives, 2007/1 Vol. XII : https://shs.cairn.info/revue-francaise-de-linguistique-appliquee-2007-1

Revue française de linguistique appliquée, Grands corpus : diversité des objectifs, variété des approches, 1999/1 Vol. IV : https://shs.cairn.info/revue-francaise-de-linguistique-appliquee-1999-1

Revue française de linguistique appliquée, Corpus. De leur constitution à leur exploitation, 1996/2 Vol. I : https://shs.cairn.info/revue-francaise-de-linguistique-appliquee-1996-2

Syntaxe & Sémantique, Textes, documents numériques, corpus. Pour une science des textes instrumentées, 2008/1 n°9 : https://shs.cairn.info/revue-syntaxe-et-semantique-2008-1

Verbum, Corpus oraux : recueil et analyse de données, 2008/4 tome XXX : https://www.atilf.fr/publications/revues-atilf/verbum/

Autres références bibliographiques

Abeillé, Anne, Godard, Danièle, Delaveau, Annie et Gautier, Antoine. 2021. *La Grande Grammaire du Français*. Actes Sud.

Aliquot-Suengas, Sophie. 2003. La productivité actuelle de la forme constructionnelle -*ade*. *Langue française*, 140, 38-55.

Aston, Guy. 2011. Applied Corpus Linguistics and the learning experience. In Vander Viana, Zyngier Sonia & Barnbrook Geoff, *Perspectives on Corpus Linguistics*, Volume 48, Studies in Corpus Linguistics, Benjamins.

Auriac-Slusarczyk, Emanuèle et Delsart, Aline. 2021. Des discours authentiques singuliers aux scénarios de formation pour les médecins : quelle méthode d'exploitation pragmatique du corpus DECLICS2016 ? *Corpus*, 22. https://journals.openedition.org/corpus/5960.

Avanzi, Mathieu, Barbet, Cécile, Glikman, Julie et Peuvergne, Julie. 2016. Présentation d'une enquête pour l'étude des régionalismes du français. *5e Congrès Mondial de Linguistique Française*. https://doi.org/10.1051/shsconf/20162703001.

Bally, Charles. 1909. *Traité de stylistique française*. Heidelberg.

Bauche, Henri. 1920. *Le langage populaire*. Paris : Payot. 4ème édition 1946.

Baude, Olivier et Dugua, Céline. 2016. Les ESLO, du portrait sonore au paysage digital. *Corpus*, 15. http://journals.openedition.org/corpus/2924.

Béguelin, Marie-José, Apothéloz, Denis, Benzitoun, Christophe, Corminboeuf, Gilles, Deulofeu, José, Lauwers, Peter et Willems, Dominique. sous presse. Décrire le français au XXIème siècle. In Neveu Franck et Fasciolo Marco (éds.). *Décrire une langue : objectifs et méthodes*. Classiques Garnier.

Bergounioux, Gabriel. 2014. Retour sur l'enquête dialectologique de Brunot en Berry et Limousin (1913). *4e Congrès Mondial de Linguistique Française*.

Berrendonner, Alain et Reichler-Beguelin, Marie-José. 1990. Décalages : les niveaux de l'analyse linguistique. *Langue française*, 81, 99-125.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4), Oxford University Press, 243-257.

Bilger, Mireille et Cappeau, Paul. 2021. *Un peu, beaucoup… pas du tout*. Le recours au quantitatif dans la linguistique sur corpus à travers l'exemple de « en mode ». In Annie Bertin, Françoise Gadet, Sabine Lehmann, Anaïs Moreno Kerdreux (éd.), *Réflexions théoriques et méthodologiques autour de données variationnelles*, Presses universitaires Savoie Mont Blanc.

Blanche-Benveniste, Claire. 1995. De la rareté de certains phénomènes syntaxiques en français parlé. *French Language Studies*, 5, 17-29.

Blanche-Benveniste, Claire. 1996. De l'utilité du corpus linguistique. *Corpus de leur constitution à leur exploitation*. *Revue Française de Linguistique Appliquée*, Vol. 1-2, 25-42.

Blanche-Benveniste, Claire. 1997. Transcription et technologie. *Recherche sur le français parlé*, 14, 87-100.

Blanche-Benveniste, Claire. 2002. Réflexions sur les transcriptions de français parlé. *Revue Parole*, 22/23/24, 91-117.

Blanche-Benveniste, Claire. 2010a. Où est le *il* de *il y a* ? *Travaux de linguistique*, 61, 137-153.

Blanche-Benveniste, Claire. 2010b. *Le français : usages de la langue parlée*. Leuven-Paris : Peeters.

Blasco, Mylène (éd). 2022. *Parler à l'hôpital. Écouter ce qui est dit, décrypter ce qui se dit*. Münster : Nodus Publikationen.

Boulton, Alex et Tyne, Henry. 2014. *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Coll. Langues et didactique. Paris : Didier.

Cappeau, Paul. 2005. Quand on ne trouve pas ce que l'on cherche… *Faits de langues*, 25, 157-160.

Cappeau, Paul (dir.). 2021. *Une grammaire à l'aune de l'oral ?* Rennes : PUR.

Cappeau, Paul et Gadet, Françoise. 2021. Transcrire c'est (déjà) analyser. *Travaux linguistiques du CerLiCO*, 31, 13-29.

Chanet, Catherine. 2001. 1700 occurrences de la particule *quoi* en français parlé contemporain : approche de la « distribution » et des fonctions en discours. *Marges Linguistiques*, 2, 56-81.

Cheng, Winnie. 2012. *Exploring Corpus Linguistics. Language in Action*. Abingdon: Routledge.

Cislaru, Georgeta et Olive, Thierry. 2018. *Le Processus de textualisation. Analyse des unités linguistiques de performance écrite*. coll. « Champs linguistiques », Louvain-la-Neuve : De Boeck.

Condamines, Anne. 2003. *Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques*. Habilitation à diriger des recherches. Université de Toulouse Le Mirail.

Condamines, Anne et Dehaut, Nathalie. 2011. Mise en œuvre des méthodes de la linguistique de corpus pour étudier les termes en situation d'innovation disciplinaire : le cas de l'exobiologie. *Meta : journal des traducteurs*. Presses Universitaires de Montréal, 56(2), 266-283. https://doi.org/10.7202/1006176ar.

Cordereix, Pascal. 2014. Ferdinand Brunot et les archives de la parole : le phonographe, la mort, la mémoire. *Revue de la BNF*, 48, 5-11.

Damourette, Jacques et Pichon, Edouard. 1911-1936. *Essai de grammaire de la langue française*. Paris : D'Artrey.

Debaisieux, Jeanne-Marie et Benzitoun, Christophe. 2020. Présentation du numéro consacré à Orféo. *Langages*, 219, 9-24.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Springer.

Doquet, Claire, Enoiu, Vanda, Fleury, Serge et Mazziotti, Sara. 2017. Problèmes posés par la transcription et l'annotation d'écrits d'élèves, *Corpus*, 16. https://doi.org/10.4000/corpus.2776.

Dostie, Gaétane. 2014. Considérations sur la forme et le sens. *Pis* en français québécois. Une simple variante de *puis* ? Un simple remplaçant de *et* ? *Journal of French Language Studies*, 14-2, 113-128.

Dostie, Gaétane. 2016. Le Corpus de français parlé au Québec (CFPQ) et la langue des conversations familières : Exemple de mise à profit des données à partir d'un examen lexico-sémantique de la séquence *je sais pas*. *Corpus*, 15. http://journals.openedition.org/corpus/2945.

Duchet, Jean-Louis, Trapateau, Nicolas et Castanier, Jérémy. 2012. Stress Placement in Pronouncing Dictionaries (1727–2010): Latin Etymology vs. English Derivation. *Language and History*. Taylor & Francis. 55(1), 34-46.

Emirkanian, Louisette et Fouqueré, Christophe. 2003. Présentation : TALN, Web et corpus. *Revue québécoise de linguistique*, 32. https://id.erudit.org/iderudit/012241ar.

Fairon, Cédric et Paumier, Sébastien. 2007. Un corpus de SMS est-il un corpus comme les autres ?. *26th conference on Lexis and Grammar*, Bonifacio. https://infolingu.univ-mlv.fr/Colloques/Bonifacio/proceedings/fairon.pdf.

Frei, Henri. 1929. *La grammaire des fautes*. Paris : P. Geuthner.

Frey, Claude et Latin, Danièle. 1997. *Le corpus lexicographique. Méthodes de constitution et de gestion*. Champs linguistiques, De Boeck Supérieur.

Gadet, Françoise. 2020. *Langue et variation*. In *Encyclopédie grammaticale du français*.

Gadet, Françoise et Guerin, Emmanuelle. 2016. Construire un corpus pour des façons de parler non standard : Multicultural Paris French. *Corpus*, 15. http://journals.openedition.org/corpus/3049.

Gary-Prieur, Marie-Noëlle. 2005. Où il est montré que le nom propre n'est (presque) jamais modifié. *Langue française*, 146, 53-66.

Gillet, Pauline. 2024. *Description syntaxique des interrogatives partielles chez les enfants francophones : situation de diglossie ou exploitations différenciées d'une unique grammaire ?* Thèse de doctorat soutenue à l'université de Lorraine.

Gougenheim, Georges, Michea, René, Rivenc, Paul et Sauvageot, Aurélien. 1964. *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris : Didier.

Gries, Stefan Th. 2021. *Statistics for Linguistics with R: A Practical Introduction* (3rd rev). Berlin: Walter de Gruyter.

Grossmann, Maria, Montermini, Fabio, Passino, Diana, Pescarini, Diego et Villoing, Florence. 2022. Présentation du numéro Corpus et données en morphologie. *Corpus*, 23. https://doi.org/10.4000/corpus.6823.

Groupe de Fribourg. 2024. *L'inversion du clitique sujet et ses fonctions en français contemporain*. Sciences pour la communication, Peter Lang.

Habert, Benoît. 2000. Des corpus représentatifs : de quoi, pour quoi, comment ? In Bilger, Mireille (éd). *Linguistique sur corpus – Études et réflexions*. Paris/Perpignan : Presses universitaires de Perpignan. 11-58.

Halliday, Michaël A.K. 1992. Language as System and Language as Instance: The Corpus as a Theoretical Construct. In Jonathan J. Webster (ed). *The Collected Works of M.A.K. Halliday*, Vol. 6 *Computational and Quantitative Studies*. London/New York: Continuum, 76-92.

Huguin, Mathilde. 2021. *Analyse morphologique des mots construits sur la base de noms de personnalités politiques*. Thèse de doctorat. Université de Lorraine.

Huguin, Mathilde, Hathout, Nabil, Lignon, Stéphanie et Namer, Fiammetta. 2024. *Pécressisme, mélenchomanie et marinolâtrie* ou comment les personnalités politiques sont (mal)traitées dans les créations dérivationnelles. In A. Aleksandrova, P. Cappeau & J.-P. Meyer (éds), *Des mots et des humains – Pour une sémantique référentielle textuelle*. Études en l'honneur de Catherine Schnedecker. Paris : L'Harmattan, 127-147.

Jacques, Marie-Paule. 2005. Pourquoi une linguistique de corpus ? In Williams, Geoffrey (éd). *La linguistique de corpus*. Rennes : Presses universitaires de Rennes.

Kilgarriff, Adam and Grefenstette, Gregory. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29-3, 335-347.

Klinger, Dominique. 2011. La grammaire pour elle-même et en elle-même… au-delà des genres ? L'exemple de La grammaire des fautes d'Henri Frei. *Linx*, 64-65, 69-84.

Kreutzer, Julia *et al*. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10, 50-72.

Landragin, Frédéric. 2021. Un corpus annoté en chaînes de référence et son exploitation : le projet Democrat. *Langages*, 224, 11-24.

Leech, Geoffrey. 2015. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer (eds). *Corpus Linguistics and the Web*. *Language and Computers*, 59, 133-149.

Léon, Jacqueline. 2005. Claimed and Unclaimed Sources of Corpus Linguistics. *Henry Sweet Society for the History of Linguistic Ideas Bulletin*, 44(1), 36-50.

Léon, Jacqueline. 2024. Saviez-vous quel a été le premier corpus informatisé ? *Le saviez-vous #17*. HTL. https://htl.cnrs.fr/le-saviez-vous-17/.

Maingueneau, Dominique. 2005. L'analyse du discours et ses frontières. *Marges Linguistiques*, 9.

Marchello-Nizia, Christine. 2014. L'importance spécifique de l' « oral représenté » pour la linguistique diachronique. In Wendy Ayres-Bennet et Thomas M. Rainsford (dir.). *L'histoire du français. État des lieux et perspectives*. Classiques Garnier.

Martinet, André. 1960. *Éléments de linguistique générale*. Paris : Armand Colin.

Martinon, Philippe. 1913. *Comment on prononce le français*. *Traité complet de prononciation pratique avec les noms propres et les mots étrangers*. Paris : Larousse.

Martinon, Philippe. 1927. *Comment on parle en français : la langue parlée correcte comparée avec la langue littéraire et la langue familière*. Paris : Larousse.

Martinot, Claire. 2005. Comment parlent les enfants de 6 ans ? Pour une linguistique de l'acquisition. Besançon : Presses universitaires de Franche-Comté.

Mayaffre, Damon. 2005. Rôle et place des corpus en linguistique : réflexions introductives. In Pascale Vergely (éd.). *Actes du colloque JETOU'2005*, Toulouse, Université de Toulouse-Le Mirail, 5-17.

Mayaffre, Damon. 2010a. Corpus et web-corpus. Réflexion sur la corporalité numérique. *Cahiers de praxématique*, 54-55. https://journals.openedition.org/praxematique/1170.

Mayaffre, Damon. 2010b. *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*. Histoire. Mémoire de synthèse. Université Nice Sophia Antipolis.

McCarthy, Michael et O'Keeffe, Anne. 2010. What are corpora and how have they evolved ? In McCarthy et O'Keeffe (eds). *The Rouledge Handbook of Corpus Linguistics*. New York: Routledge, 3-13.

Mellet, Sylvie. 2002. Corpus et recherches linguistiques : introduction. *Corpus*, 1. https://doi.org/10.4000/corpus.7.

Mondada, Lorenza. 2001. Pour une linguistique interactionnelle. *Marges Linguistiques*, 1.

Mondada, Lorenza. 2017. Nouveaux défis pour l'analyse conversationnelle : l'organisation située et systématique de l'interaction sociale. *Langage et société*, 160-161, 181-197.

Muller, Charles. 1993. *Initiation aux méthodes de la statistique linguistique*. Paris : Honoré Champion.

Muni Toke, Valelia. 2012. Le linguiste et le médecin. *Genesis*, 35, 101-108. https://doi.org/10.4000/genesis.1054.

Muni Toke, Valelia. 2013. *La grammaire nationale selon Damourette et Pichon 1911-1939*. Paris: ENS Éditions.

Oakey, David. 2009. Fixed Collocational Patterns in Isolexical and Isotextual Versions of a Corpus. In Paul Baker (ed). *Contemporary Corpus Linguistics*. London/New-York: Continuum, 140-158.

Paillé, Pierre et Mucchielli, Alex. 2012. Chapitre 1 - Choisir une approche d'analyse qualitative, *L'analyse qualitative en sciences humaines et sociales.* sous la direction de Paillé Pierre, Mucchielli Alex. Armand Colin, pp. 13-32.

Paveau, Anne-Marie. 2014. L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques. *Corela*, HS-15.

Pernier, Maurice. 1986. *Le mot*. Paris : PUF.

Pincemin, Bénédicte. 2012. Hétérogénéité des corpus et textométrie. *Langages*, 187, 13-26. https://www.cairn.info/revue-langages-2012-3-page-13.htm.

Rastier, François. 2011. *La mesure et le grain*. Paris : Honoré Champion.

Reppen, Randi. 2011. Building a corpus - What are the key considerations? In Anne O'Keeffe et Michael McCarthy (eds). *The Routledge Handbook of Corpus Linguistics*. London / New York: Routledge, 31-37.

Rey, Alain (dir). 1992. *Dictionnaire historique de la langue française*. Paris : Le Robert.

Rouget, Christine. 2005. Exception et linguistique sur corpus. *Faits de langues*, 25, 151-155.

Scheer, Tobias. 2004. Présentation du volume. En quoi la phonologie est vraiment différente. Revue *Corpus*, 3. http://journals.openedition.org/corpus/193.

Scheer, Tobias. 2013. The Corpus: A Tool among Others. *Corela*, 13. http://journals.openedition.org/corela/3006.

Sctrick, Robert. 1968-. CORPUS, *linguistique*, *Encyclopædia Universalis* [en ligne], consulté le 20 avril 2023. URL : http://www.universalis-edu.com/encyclopedie/corpus-linguistique/.

Schmidt-Lainé, Claudine et Pavé, Alain. 2008. La modélisation au cœur de la démarche scientifique et à la confluence des disciplines. *Les Cahiers du Musée des Confluences*. Revue thématique Sciences et Sociétés du Musée des Confluences, tome 2, L'Expérimentation, 21-34.

Tancoigne, Élise, Corbellini, Jean-Philippe, Deletraz, Gaëlle, Gayraud, Laure, Ollinger, Sandrine et Valero, Daniel. 2020. *La transcription automatique : un rêve enfin accessible ? Analyse et comparaison d'outils pour les SHS*. Nouvelle méthodologie et résultats. https://hal.science/halshs-02917916v2.

Thiberge, Gabriel. 2020. *Acquisition et maîtrise des interrogatives partielles en français : La variation sociolinguistique comme outil interactionnel*. Thèse de doctorat. Université de Paris.

Thuilier, Juliette. 2012. *Contraintes préférentielles et ordre des mots en français*. Thèse de doctorat. Université Paris 7.

Vaguer, Céline. 2007. Corpus, vous avez dit corpus ! de la notion de corpus à la création d'un « corpus informatisé ». In Williams, Geoffrey (éd). *Corpus, Langues et Linguistique*, 207-223.

Valli, André et Véronis, Jean. 1999. Étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue française de linguistique appliquée*, Vol. IV-2, 113-133.

Vergne, Jacques et Giguet, Emmanuel. 1998. Regards Théoriques sur le "Tagging". *Actes de la conférence TALN*. 22-31.

Williams, Geoffrey. 2006. La linguistique de corpus : une affaire prépositionnelle. *Texto*. http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf.

Corpus en langue française cités

Nom du corpus	Adresse	Type d'accès
88milSMS	http://88milsms.huma-num.fr/	sur demande
Araeneum Francogallicum III	http://unesco.uniba.sk/	sur inscription
CEFC Corpus Évolutif du Français Contemporain	https://www.ortolang.fr/market/corpora/cefc-orfeo https://orfeo.ortolang.fr/	Accès libre
CFPP2000 Corpus de Français Parlé Parisien des années 2000	http://cfpp2000.univ-paris3.fr	Accès libre sur le site, sur CoCoON et sur Ortolang. Recherche de mots / séquences de mots
CFPQ Corpus de français parlé au Québec	https://applis.flsh.usherbrooke.ca/cfpq/	Accès libre mais sans l'audio ni la vidéo Recherche de mots / expressions

CLAPI Corpus de Langue Parlée en Interaction – Banque de données multimédia	http://clapi.ish-lyon.cnrs.fr	Accès libre à 67h d'enregistrements
Corpus 14	https://www.univ-montp3.fr/corpus14/	Accès libre via le portail TXM
DECLICS2016		Sur demande
E-CALM	https://www.ortolang.fr/market/corpora/e-calm/v2.1	Accès libre
Ecriscol écrits produits en situation scolaire	http://www.univ-paris3.fr/ecriscol	Accès libre
ESLO Enquêtes sociolinguistiques à Orléans	http://eslo.huma-num.fr	Accès libre Recherche de mots
FLEURON Français Langue Étrangère Universitaire Ressources et Outils Numériques	https://fleuron.atilf.fr	Accès libre Recherche de mots
FLORALE Français langue orale pour le FLE	https://florale.unil.ch/	Accès libre – recherche
Frantext Textes du 9ème au 21ème siècles	http://www.frantext.fr	Inscription payante
French Treebank	http://ftb.linguist.univ-paris-diderot.fr/	sur demande
FrWaC	https://www.clarin.si/noske/run.cgi/corp_info?corpname=frwac&struct_attr_stats=1	Accès libre avec l'outil nosketch engine
Grand corpus des dictionnaires	https://classiques-garnier.com/?view=folder&lang=fr_FR&folder_id=27	Accès restreint
Grand corpus des grammaires françaises, des remarques et des traités sur la langue	https://classiques-garnier.com/grand-corpus-des-grammaires-fran%C3%A7aises-des-remarques-et-des-trait%C3%A9s-sur-la-langue-xive-xviiie-s.html	Accès restreint
MPF	https://www.ortolang.fr/market/corpora/mpf	Se connecter à Ortolang pour

		accéder au corpus
OFROM corpus Oral de Français de Suisse Romande	http://ofrom.unine.ch/	Accès libre sur le site, sur NAKALA et CoCoON – Différents types de recherche (mots, lemmes, séquences)
PFC Phonologie du français contemporain	https://www.projet-pfc.net	sur inscription
Prize papers	https://www.prizepapers.de/	Pas d'accès – projet en cours
Scientext	https://corpora.aiakide.net/scientext20/	Accès libre - recherche
Scoledit dictées et productions écrites d'élèves	http://scoledit.org/scoledition/	Accès libre
TCOF Traitement des corpus oraux du français	https://tcof.atilf.fr/	Accès libre

Il existe également le Fonds de données linguistiques du Québec (https://fdlq.recherche.usherbrooke.ca), une plateforme regroupant divers corpus oraux, écrits, métalinguistiques et dialectologiques. Il s'agit d'une riche base de données dont la plupart des corpus ont malheureusement un accès limité.