Autour du corpus exploité dans l'étude du discours scientifique

Around the corpus used in the study of scientific discourse

Noureddine NAJAI*,
Laboratoire *Langues, Discours et Cultures*,
ISSHJ (Tunisie),
benmohamednour@hotmail.fr

Date de soumission : **20.03.2023**
Date d'acceptation : **05.04.2024**
Date de publication : **10.04.2024**

Ex PROFESSO

Volume 09 / Numéro 01 / Année 2024

Résumé

S'il est une unanimité parmi les linguistes concernant le corpus, c'est la nécessité dans tout travail de recherche de définir cette notion, car c'est le corpus qui amène le chercheur à pouvoir formuler une hypothèse, à en éprouver la consistance ou, au contraire, à en montrer les limites. Mais cette tâche n'est pas aussi simple que l'on pourrait penser vu les problèmes qu'elle peut poser.

Le présent article constitue une réflexion sur les difficultés que peut rencontrer le linguiste travaillant sur le discours scientifique. Autant de questions inquiétantes pour le chercheur en rapport avec le corpus seront exposées ici et auxquelles nous tenterons d'apporter des réponses : selon quels critères constitue-t-on son corpus ? Convient-il d'opter pour une approche quantitative ou pour une approche qualitative, ou encore selon un modus vivendi entre les deux ? Pourquoi privilégie-t-on le genre de l'article aux dépens des autres genres ? etc.

Mots-clés : corpus ; difficultés ; discours scientifique ; article de recherche ; fluctuations ; démarche ;

Abstract

If there is unanimity among linguists concerning the corpus, it is the necessity in any research work to define this notion, because it is the corpus that leads the researcher to be able to formulate a hypothesis, to test its consistency or, on the contrary, to show its limits. But this task is not as simple as one might think given the problems it can pose.

This article is a reflection on the difficulties that the linguist working on scientific discourse may encounter. So many worrying questions for the researcher in relation to the corpus will be exposed here and to which we will try to provide answers: according to which criteria do we constitute his corpus? Should we opt for a quantitative approach or a qualitative approach, or a modus vivendi between the two? Why is the genre of the article preferred over other genres? etc.

Keywords : corpus ; difficulties ; scientific discourse ; research article ; approach.

Url de la revue :
https://www.asjp.cerist.dz/en/PresentationRevue/484

* - *Auteur correspondant.*

INTRODUCTION

La question des corpus est cruciale dans le travail ressortissant aux sciences du langage en général, si bien qu'aucun chercheur ne peut s'en passer. Celui-ci ne peut, en effet, traiter d'un phénomène linguistique donné sans être attentif à cette opération de collecte de données et aux questions qu'elle charrie, à savoir la manière de faire la collecte, l'outil utilisé dans le traitement des données, la démarche adoptée, etc. De fait, cette tâche n'est pas sans difficultés, notamment en présence de la controverse conceptuelle, terminologique, méthodologique, etc. qui tourne autour de la notion de *corpus* dans la littérature.

Nous nous focaliserons dans le présent article précisément sur les difficultés que pourrait rencontrer le chercheur/linguiste travaillant sur le discours scientifique lors de la constitution et l'analyse de son corpus. Nous tenterons de répondre à quelques questions qui pourraient l'inquiéter telles que, selon quels critères devrait-on constituer son corpus ? Quelles fluctuations devrait-on prendre en compte dans cette opération pour que les résultats soient bien fondés et généralisables ? Conviendrait-il d'opter pour une approche quantitative ou pour une approche qualitative, ou encore selon un modus vivendi entre les deux ? Pourquoi devrait-on privilégier le genre de l'article aux dépens des autres genres ? etc. Notre objectif principal est avant tout de fournir, en partant de notre modeste expérience en la matière, quelques repères qui nous semblent utiles dans le choix et l'analyse linguistique d'un corpus d'écrits scientifiques. Nous ne prétendons pas proposer une recette à suivre, mais nous tenterons d'éclairer quelques pistes susceptibles, selon nous, d'aider le chercheur à exécuter cette tâche.

Dans un premier temps, nous ferons un bilan non exhaustif de quelques écueils auxquels se heurte le chercheur en sciences du langage en général dans l'établissement de son corpus d'analyse. Puis, nous ferons le tour des spécificités du corpus en analyse du discours. Ce faisant, nous présenterons les différentes conceptions que se font les analystes du corpus, ainsi que les diverses approches qu'ils adoptent pour l'analyser. Ensuite, nous pointerons les principaux critères qui s'imposent au chercheur traitant du discours scientifique, aussi bien dans la constitution que dans l'analyse de son corpus. Nous tenterons de montrer à travers quelques exemples l'importance cruciale de ces critères. Nous terminerons par un bref aperçu sur *Scientext*, un vaste corpus d'écrits scientifiques qui pourrait être utile pour le chercheur souhaitant mener une étude linguistique sur ce type de discours, et ce pour différentes raisons.

I. CONSTITUTION DE CORPUS EN SCIENCES DU LANGAGE. UNE TÂCHE NON SANS ÉCUEILS

En sciences du langage, le recueil des données est une tâche non sans écueils. En effet, si le chercheur arrive plus au moins aisément à choisir la matérialité langagière de ces données (productions orales ou productions écrites) et le support qui les véhicule, d'autres points, qui n'ont toujours pas donné lieu à un consensus, lui posent pourtant des difficultés. Nous nous contenterons d'en citer, à la suite de Charaudeau (2009), quelques-uns.

Il y a d'abord le problème de la représentativité des données recueillies qui a été déjà longuement discuté dans la littérature sur le corpus. Se pose ici la question si un corpus permet ou non de déboucher sur une sorte de généralisation ou modélisation.

Dans cette perspective, Mellet (2002 : 6) estime qu'« 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 pouvoir prétendre à être représentatif d'autre chose que de lui-même ». De surcroît, un échantillon de langue que forme le corpus - car un corpus ne saurait refléter la langue dans sa totalité – peut laisser passer sous silence des traits importants de la langue et insister *a contrario* sur des traits ordinaires (Gleason, 1969) [1].

Dans le même ordre d'idées, Vaguer (2007 : 207) considère qu'un corpus, vu qu'il représente un texte produit par un ou des locuteurs particuliers dans un temps et un espace précis et selon une certaine intention, ne peut pas rendre compte de tous les cas de figure du phénomène linguistique étudié. Ce problème charrie en fait un autre qui est celui des variables, notamment de nature extralinguistique, « telles que les types de locuteurs, les dispositifs de communication, de même les variables concernant le temps (historicité) et l'espace (les cultures) » Charaudeau (2009 : 38).

Notons, en outre, qu'il existe deux façons pour le chercheur en sciences du langage de constituer un corpus : soit il opte pour des « exemples attestés », c'est-à-dire des données préalables à sa recherche, puisées dans des textes d'auteurs (romans, revues, articles de presse, etc.), soit il produit lui-même des exemples pour les besoins de la cause, lesquels sont ainsi dits « exemples forgés »[2].

L'une ou l'autre de ces deux démarches a des avantages et des inconvénients. Si ce n'est pas le lieu ici de tenter un inventaire détaillé et exhaustif de ces avantages et inconvénients, il convient quand même d'en rappeler les plus saillants[3]. Le corpus forgé, par exemple, se prête le mieux aux différentes manipulations faites par le linguiste (commutation, clivage, extraction, déplacement, inversion, etc.) dans son travail d'analyse. Cette éventualité est peu probable pour le corpus attesté, qui, de par la longueur et la complexité de ses phrases[4], peut poser des difficultés de manipulation. L'inconvénient du corpus forgé est qu'il n'est pas réfractaire à des jugements d'acceptabilité et de grammaticalité erronés. Le risque d'erreurs provient de l'influence, voire de « la prégnance de l'hypothèse que l'on a en tête » (Vaguer, 2007 : 211).

Par contre, les données attestées, vu qu'elles sont produites extérieurement au linguiste et à son travail de recherche, ne présentent normalement pas ce risque d'être erronées sous l'influence de l'hypothèse formulée. Mais, quelque spontané que soit le corpus attesté, il ne garantit pas l'objectivité du linguiste. En effet, dans l'opération de sélection des données et leur mise en ordre, celui-ci se laisse guider également par une chose à découvrir, pour reprendre les termes de Vaguer (2007 : 212), et ne choisit par conséquent que les exemples compatibles avec ses hypothèses.

II. PARTICULARITÉS DU CORPUS EN ANALYSE DU DISCOURS

Il convient de prime abord de signaler qu'en analyse du discours la question du corpus, de sa constitution, de son traitement n'est pas récente. Elle a été, en effet, soulevée à l'origine par Dubois (1969), en particulier dans le numéro 13 de *Langages* fondateur de cette discipline en France. Et depuis, autant de travaux ont été consacrés à ce sujet au point qu'« aucune étude qui entend appréhender les discours ou les textes dans leur dimension sociale ou politique n'est indifférente à la question des corpus » (Mayaffre, 2005 : 1).

II.1. Productions langagières en situation d'usage

Le corpus en analyse du discours est fait de productions langagières en situation d'usage, c'est-à-dire de données empiriques, réellement produites par des sujets inscrits dans un contexte particulier. Si l'on se reporte à la dichotomie saussurienne langue / parole (reformulée ensuite en langue / discours), le corpus fera ici l'objet d'analyse de ce que Charaudeau (2009 : 41) appelle « linguistique du discours », qui, par opposition à « linguistique de la langue », ne se contente pas de décrire les systèmes intrinsèques à la langue sans prendre en compte les usages et les significations sociales.

On voit bien que l'objet d'analyse en analyse du discours est le texte dans sa relation avec son contexte de production. Là, il convient de distinguer *texte* et *discours* car le texte n'est pas le discours et le discours n'est pas le texte. Charaudeau (2009 : 40) oppose de façon nette ces deux notions et il critique d'ailleurs les ouvrages qui, bien qu'ils se réclament de l' « analyse textuelle », emploient les termes « analyse du discours », et *vice versa*. Le texte, comme le dit bien Adam (1997), se définit selon un critère de cohésion, c'est-à-dire par l'organisation de sa configuration en rapport avec ce qui l'entoure (le cotexte), et le discours selon un critère de cohérence qui fait intervenir, outre la cohésion, le contexte[5] et le genre de discours. Selon Charaudeau (2009 : 44), « le discours est un parcours de signifiance qui se trouve inscrit dans un texte, et qui dépend de ses conditions de production et des locuteurs qui le produisent et l'interprètent ». Le texte est donc une sorte de réceptacle en ce qu'il est porteur de discours. Un texte peut porter un ou plusieurs discours et un discours peut « irriguer des textes différents ».

II.2. Valeur heuristique *vs* valeur validante

Le corpus a une valeur heuristique ou validante selon que l'approche adoptée par l'analyste est inductive ou hypothético-déductive.

Dans le premier cas de figure, la priorité est donnée aux données, à partir desquelles on peut générer des théories. « C'est finalement le corpus qui fait la théorie », le dit bien Dalbera (2002 : 9) plaçant ainsi le corpus au centre de la démarche inductive. Dans cette optique, Mayaffre (2005 : 3) considère qu' « un corpus n'est pas un réceptacle mais une matrice » en ce sens qu'il est un lieu d'invention, « un processus créateur » (Mellet, 2002 : 9).

Dans le second cas de figure, la priorité est donnée à l'hypothèse et à la mise en place du cadre théorique et c'est cette hypothèse qui préside à la constitution du corpus, puis l'analyse permet ensuite de la confirmer ou l'infirmer. Cette démarche hypothético-déductive, contrairement à celle inductive, se propose de déboucher sur « des résultats prédictibles, vérifiables et généralisables » (Normand, 2014 : 12), et c'est cette démarche qu'adoptent la plupart des chercheurs.

Notons que d'aucuns, tout en pensant que la constitution d'un corpus dépend d'un positionnement théorique non sans lien avec un objectif d'analyse *a priori*, ne nient pourtant pas le fait que le corpus participe également d'une démarche heuristique. C'est, entre autres, le cas de Charaudeau (2009 : 55-56), qui affirme qu'en sciences humaines et sociales de manière générale la démarche est double, en ce sens que l'une ne va pas sans l'autre.

II.3. Analyse quantitative *vs* analyse qualitative

Les analyses de corpus en analyse du discours oscillent entre deux orientations, l'une privilégie l'approche qualitative et l'autre l'approche quantitative. S'il y a un consensus sur le fait que c'est la taille du corpus qui impose l'une ou l'autre de ces deux approches[6], les chercheurs s'inscrivant dans une optique qualitative reprochent à ceux se situant dans une optique quantitative, entre autres, de se limiter à la stricte matérialité textuelle. De même, ceux-ci reprochent aux tenants de l'approche qualitative de « se fonder sur des intuitions, d'aboutir à une herméneutique trop subjective sans tenir compte de la relativité des phénomènes au sein d'un corpus donné » (Rouveyrol, 2005 : 1).

En dépit de ce débat, il y a actuellement une tendance à concilier les deux méthodes. L'analyse quantitative ne peut, à elle seule, assurer la fiabilité de l'analyse. Certes, elle a du sens en soi, mais il s'agit d'un « sens provisoire devant être confirmé, corrigé, voire contredit, et en tout cas étendu et approfondi par l'analyse qualitative » (Charaudeau, 2009 : 64). L'engouement que les chercheurs ont aujourd'hui pour l'approche quantitative ne doit pas neutraliser l'approche qualitative, sans quoi le pouvoir explicatif de l'analyse du discours serait réduit.

III. LA NOTION DE CORPUS DANS L'ANALYSE DU DISCOURS SCIENTIFIQUE

Avant d'appréhender la notion de *corpus* dans l'analyse du discours scientifique (désormais DS), il nous semble utile de définir succinctement ce type de discours.

III.1. Qu'est-ce que *le discours scientifique* ?

Le DS sera entendu ici comme « discours produit dans le cadre de l'activité de recherche à des fins de construction et de diffusion du savoir » (Rinck, 2010 : 428). Le qualificatif « scientifique » selon cette optique ne se cantonne pas aux sciences *stricto sensu*, c'est-à-dire les sciences de la nature, il se rattache en outre à toute activité académique quelle que soit la discipline dont elle fait partie. Le DS comporte des genres écrits (monographies, articles, HDR[7], thèses, etc.) et des genres oraux (communautés de conférences, séminaires, etc.) se rapportant aux sciences dites « dures » ou à celles dites « douces »[8]. Rappelons au passage qu'en contexte anglo-saxon, on parle de « *academic discourse* », car « *scientific discourse* » ne désigne que les disciplines des sciences dures.

Les partitions entre les différents domaines scientifiques et les disparités qui les marquent au niveau des objets, des méthodes, des enjeux de connaissance, des conventions rédactionnelles, etc. conduisent certains linguistes, pour ne citer que Grossmann (2012), à privilégier plutôt le pluriel du terme « discours scientifique ». Les variations d'ordre culturel, disciplinaire, temporel, graduel, individuel, etc. qui marquent d'ailleurs les textes scientifiques, ainsi que l'absence d'un style scientifique universel, font que l'unicité tant proclamée du DS n'est pas un constat pertinent.

En analyse du discours, le DS est considéré comme un discours « fermé » (Charaudeau et Maingueneau, 2002 : 261) car son auteur en est à la fois l'origine et la cible. En d'autres termes, c'est un discours émanant d'un locuteur spécialisé et s'adressant à un allocutaire, lui aussi, spécialisé dans le même domaine scientifique. On parle de fait de communauté scientifique, qui est une communauté restreinte aux producteurs et aux récepteurs de ce discours, ce qui n'est pas le cas, par exemple, du

discours politique, qui est un discours ouvert car il s'adresse à un public vaste et varié, en ce sens qu'il n'a pas nécessairement la même caractérisation sociale.

Il existe toutefois une tendance s'appuyant sur la désignation « *academic discourse* », qui considère que la transposition didactique du discours des chercheurs à l'université et les écrits des étudiants dans leur parcours universitaire et leur formation à la recherche relèvent également du DS. Rinck (2010 : 429) place le DS dans un cadre plus large qu'elle appelle « la société de la connaissance », et ce à travers ses reprises ou sa circulation dans les autres types de discours, à savoir les discours universitaire, médiatique, ou de l'expertise.

III.2. Le genre de l'article

Certes, l'étude du DS peut se faire à partir de différents genres, aussi bien écrits qu'oraux, mais c'est le genre de l'article qui est le plus recommandé dans les travaux portant sur ce type de discours. L'attention particulière qu'on accorde souvent à l'article aux dépens des autres genres peut s'expliquer par le fait que ce genre est le plus accrédité, si bien qu'il acquiert un statut emblématique dans l'activité scientifique. Cette accréditation passe surtout par la soumission de l'article aux différentes instances concernant sa recevabilité, aussi bien au sein du laboratoire dans les conversations informelles entre les chercheurs ou lors du contrôle et de l'évaluation du directeur du laboratoire, que dans ce va-et-vient entre l'auteur de l'article et les responsables de contrôle de la publication. L'accréditation de l'article, de son contenu et, d'une manière indirecte, de son producteur est inhérente également à sa conformité aux exigences d'ordre formel imposées par les instances de publication : la feuille de style de la revue, sa présentation, sa longueur et la place des encarts non-textuels, le style de bibliographie, etc. Les étapes par lesquelles passe l'article pour être enfin validé représentent à la fois un gage de scientificité et une reconnaissance académique du chercheur dans la communauté scientifique. L'article scientifique obéit à des règles et à des codes tant au niveau du contenu qu'au niveau de la forme (Boure, 1998 : 107), ce qui en fait un genre structuré auquel se fient les chercheurs dans l'échange des résultats de recherches scientifiques. De plus, en comparaison avec les autres genres relevant notamment de l'écrit, comme les monographies, les thèses et les HDR, l'article est plus accessible et il se prête le mieux, de par sa longueur relativement limitée, à l'étude.

Ce sont surtout ces raisons qui font que le genre de l'article est le corpus privilégié des recherches sur le DS. Il est toutefois possible de travailler sur un corpus fait de textes appartenant à des genres différents, et ce dans une perspective comparative. Tutin *et al.*(2009), pour ne citer que cet exemple, ont opté pour ce choix et abouti à la conclusion qu'il y a plus de marques d'opinion dans les mémoires d'HDR que dans les autres genres.

III.3. Fluctuations à prendre en compte

Une fois la question du genre d'écrits est résolue[9], le chercheur travaillant sur le DS se heurte dans la constitution de son corpus à une multitude de choix dus à plusieurs types de fluctuations. Il se trouve en effet contraint de prendre en compte dans l'exécution de cette tâche des paramètres additionnels, à savoir le temps (historicité), la culture, le domaine scientifique auxquels se rattachent les articles objet de l'analyse.

Les articles scientifiques relevant des sciences de la nature, par exemple, obéissent à des conventions et pratiques rédactionnelles standardisées telles que le format IMRED (acronyme pour Introduction, Matériel et Méthodes, Résultats et Discussion, connu en anglais sous la forme IMRAD), ce qui n'est généralement pas le cas pour les sciences humaines et sociales comme la linguistique, qui, bien qu'elle voie ce format gagner du terrain, ses textes suivent très souvent une organisation d'ordre thématique. En d'autres termes, les sous-titres sont plutôt régis par le contenu véhiculé dans les sections que par les fonctions que remplissent celles-ci (Rinck, 2006 ; Poudat, 2006).

Il existe également des fluctuations d'ordre intra-disciplinaire qui se manifestent, entre autres, dans la fragmentation en sous-disciplines. C'est le cas, par exemple, de la psychologie, qui se scinde en psychologie clinique et psychologie cognitive. Ces fluctuations traversant l'intérieur des disciplines apparaissent également dans les démarches et les outils utilisés. Prenons l'exemple de la phonétique actuelle, qui, bien qu'elle appartienne aux sciences humaines, se veut expérimentale. Il en est de même pour la médecine, qui, tout en privilégiant les données expérimentales, comporte aussi « une sémiologie des symptômes et un « savoir-faire » qui la rapproche dans certains cas de l'herméneutique et même d'un « art », etc. » (Grossmann, 2012 : 144).

Il conviendrait également, pour le chercheur, de prendre en compte dans la constitution du corpus les variations culturelles. Les Français, par exemple, ont une conception idéalisée du DS, profondément ancrée dans la tradition aristotélicienne. Il s'agit en fait d'une conception fidèle à cette image déshumanisée de la science défendue par la logique[10], d'où l'abandon du *moi* au profit d'une figure impersonnelle rendue par le pluriel de modestie *nous*, le *on* générique, l'hortatif, les structures impersonnelles, le passif (surtout non agentif), etc.

La culture anglo-saxonne, quant à elle, privilégie la présence explicite de l'auteur dans ses écrits à travers, entre autres, le recours à la première personne du singulier. Cette présence est appréhendée en tant que signe de sincérité, voire une façon d'ouvrir la discussion avec le lecteur, qui, selon cette conception, est plutôt un partenaire scientifique qu'un simple récepteur inconditionné (Reutner, 2010 : 82). Selon certains analystes (voir, entre autres, Kaufer et Geisler, 1989), cette tendance à personnaliser le discours s'explique par l'impact de l'individualisme et de la compétitivité de la société d'aujourd'hui, qui font que l'auteur s'efforce de manifester sa propre originalité et de laisser ses empreintes personnelles dans un domaine scientifique donné.

Un autre type de fluctuations devrait être aussi pris en compte au même titre que les autres fluctuations dans l'établissement et l'analyse du corpus : il s'agit du statut de l'auteur du genre étudié. En effet, il ne conviendrait pas de mettre sur le même plan un écrit scientifique dont l'auteur est un chercheur chevronné et un écrit scientifique appartenant à un chercheur néophyte. On ne peut pas, par exemple, appréhender de la même manière une thèse, dont l'auteur est un chercheur débutant ayant comme ambition première se faire une place dans la communauté scientifique, et un HDR, dont l'auteur est au contraire un chercheur confirmée aspirant à diriger ou à codiriger une thèse ou à siéger comme rapporteur de thèse. Nous renvoyons ici à titre d'exemple à Boch et Grossmann (2002) qui, dans une étude comparative entre experts et novices quant à la référence au discours d'autrui, constatent entre autres que ces derniers n'utilisent ce discours que pour donner une définition, introduire un

propos ou appuyer une affirmation, et non pour marquer un positionnement, comme, par exemple, souligner son adhésion à une école ou un courant de pensée, ou au contraire sa démarcation d'une position exprimée par un auteur.

Pour résumer, l'analyste du DS est appelé à respecter autant que faire se peut l'ensemble des critères indiqués *supra* pour que son étude, qu'elle soit synchronique ou diachronique, aboutisse à des résultats pertinents. Toutefois, l'homogénéité du corpus demeure difficile à atteindre, notamment en présence de la variation interne qui apparaît dans les marges que certains chercheurs se permettent par rapport aux normes régissant le même genre, le même domaine scientifique, la même (sous) discipline, la même culture, etc.

III.4. Modus vivendi entre le quantitatif et le qualitatif

L'apogée du traitement automatique des langues, de la statistique et plus généralement de la linguistique du corpus a développé un engouement chez les analystes pour les analyses quantitatives. Mais, comme nous l'avons déjà indiqué plus haut, ces analyses doivent être enrichies et approfondies par des analyses qualitatives.

Prenons un exemple. Dans une étude sur la présence pronominale du chercheur dans les écrits scientifiques, il ne suffit pas de compter les occurrences des pronoms personnels sujets en présence, à savoir *je*, *nous* et *on*. Il conviendrait également d'opter pour une étude qualitative du corpus prenant en compte le contexte, notamment verbal, pour pouvoir déceler les valeurs sémantico-référentielles de ces pronoms. Le *je*, abstraction faite de cas où il est question de discours rapportés ou de citations, ne renvoie qu'à l'auteur. Les pronoms *nous* et *on*, quant à eux, se caractérisent par une certaine ambigüité référentielle, en ce sens qu'ils ont un référent qui n'est pas stable et peut changer au cours du texte. Ce brouillage énonciatif, pour reprendre les termes de Tutin (2010 : 21), peut faire partie d'une stratégie délibérée de la part de l'auteur, car le DS est un discours rhétorique ayant des aspects persuasif et interactionnel aussi bien qu'informatif.

Hormis leur emploi générique, ces pronoms peuvent aussi bien renvoyer à un auteur collectif (un ensemble d'auteurs) ou singulier (*nous* et *on* de modestie) qu'inclure d'autres instances énonciatives, comme le lecteur (emploi dit « inclusif ») ou la communauté de discours[11] (emploi dit « exclusif »).

Il importerait également de pointer, dans une perspective exploratoire, les patrons *je-nous-on* + Verbe pour pouvoir rattacher cette présence à la fonction auctoriale qu'elle remplit. Les pronoms personnels référant à l'auteur d'une manière ou d'une autre renvoient, selon les verbes auxquels ils sont associés, à une fonction auctoriale particulière. Fløttum et Vold (2010 : 44-45) parlent de trois fonctions : *scripteur*, *chercheur* et *argumentateur*. La fonction de *scripteur* renvoie à la structuration textuelle de l'écrit, et se réalise à travers la combinaison des pronoms en question avec un verbe de discours (comme *discuter*, *illustrer*, *présenter*), suivi généralement d'une expression métatextuelle (du type *ici*, *dans ce qui suit*). Quant à la fonction de *chercheur*, elle réside dans l'activité de recherche *stricto sensu*, les verbes utilisés sont dans ce cas des verbes qui renvoient à cette activité (comme *analyser*, *considérer*, *comparer*). Pour ce qui est de la fonction d'*argumentateur*, elle concerne le positionnement de l'auteur à travers le recours à des verbes d'opinion (comme *affirmer*, *contester*, *soutenir*) et à des modalisateurs épistémiques (comme *sans doute*, *probablement*, *certainement*). A ces trois fonctions s'ajoute la fonction d'*évaluateur* (Fløtum et *al.*, 2008 : 87), qui est une fonction rarement remplie par l'auteur. Elle

apparaît dans l'emploi de modalisateurs axiologiques, comme *intéressant*, *utile*, *étonnant*, etc.

On voit bien, à partir de ces exemples portant sur la présence pronominale du chercheur dans son texte, que l'approche quantitative ne peut pas, à elle seule, assurer une analyse linguistique fiable et pertinente du DS. Se cantonner au niveau quantitatif serait attribuer à tort aux pronoms en question les mêmes valeurs sémantico-référentielles, négliger la valeur du cotexte, laisser passer sous silence la ou les fonction(s) auctoriale(s)[12] que chacun de ces pronoms remplit, etc. Le recours à l'approche qualitative du corpus s'avère donc indispensable pour vérifier et approfondir les résultats obtenus suite aux analyses quantitatives.

III.5. *Scientext*, un corpus déjà là

***S**cientext* est un vaste corpus électronique d'écrits scientifiques (thèses, articles, HDR, etc.) relevant de 3 familles de disciplines, à savoir les sciences humaines (linguistique, psychologie, sciences de l'éducation, TAL), les sciences expérimentales (biologie, médecine) et les sciences appliquées ou sciences pour l'ingénieur (électronique, mécanique). Il est gratuitement interrogeable en ligne sous sa version dédiée au grand public[13], qui compte en gros 7 HDR, 45 articles, 112 communications et 41 thèses. Les auteurs de ce projet estiment que ce corpus a été conçu pour être représentatif des différents genres et disciplines scientifiques (voir Tutin et Grossmann 2014 : 19). Il était impossible pour eux d'introduire la totalité des disciplines scientifiques, mais ils se sont cantonnés aux disciplines qui paraissent, à leurs yeux, représentatives de familles scientifiques plus larges. *Scientext* est, à notre connaissance, le seul corpus électronique offrant des écrits scientifiques français et anglais avec des outils permettant des requêtes variées sur ces écrits. Le linguiste pourrait s'en servir dans les études linguistiques portant sur le positionnement de l'auteur, les enjeux de connaissance du DS, ses routines phraséologiques, etc., et ce à partir des marqueurs lexicaux, grammaticaux et énonciatifs.

Nous pensons que cette base de données serait utile pour le chercheur/linguiste traitant du DS pour différentes raisons. D'abord, elle lui permettrait de gagner du temps et le dispenserait de l'effort voire de la peine de cette opération de collecte et de traitement des données. Ensuite, étant librement interrogeable, elle permettrait aux autres chercheurs exploitant le même domaine de vérifier ses propos et aussi de compléter ses analyses à partir du même corpus. De plus, en exploitant un corpus déjà là, c'est-à-dire fait extérieurement à sa recherche, le linguiste éviterait le risque de s'imprégner de l'hypothèse qu'il a en tête, ce qui pourrait conférer à son travail et aux résultats atteints un certain degré d'objectivité. Cette base de données offre également la possibilité de pointer une discipline scientifique bien déterminée, de travailler sur un genre bien précis puisque, comme nous venons de le signaler plus haut, chaque genre d'écrits a ses particularités, lesquelles particularités sont inhérentes au statut du chercheur, au contexte de la recherche menée, à sa finalité, aux normes rédactionnelles exigées, etc. Il est également possible de sélectionner l'une des parties principales des textes proposés (introduction, développement, conclusion) ou l'une de ses autres parties (résumé, notes de bas de page, remerciements, annexes, avant-propos, mots-clés). Cela nous semble très important pour les chercheurs qui souhaiteraient travailler sur une partie textuelle bien déterminée car, comme on le sait, chaque partie a ses spécificités. L'introduction et la conclusion, par exemple, se distinguent par leur aspect un peu technique, particulièrement propices aux marques linguistiques indiquant le positionnement du

chercheur. Dans l'introduction, ce dernier cherche non seulement à justifier l'intérêt de l'étude menée, mais aussi à se situer par rapport à ses devanciers, entre autres par le choix de la tradition ou du cadre théorique dans lequel il s'inscrit. Dans une étude sur les verbes de positionnement dans les écrits scientifiques, Tutin (2010 : 17) observe une nette surreprésentation du lexique évaluatif dans les introductions et les conclusions par rapport aux autres parties textuelles. Notons également que cette base de données a ceci de particulier qu'elle faciliterait les études contrastives par disciplines et par langues.

Nous nous contenterons de reproduire dans la partie « Annesxes », sous forme de tableau, quelques informations sur le corpus d'écrits scientifiques français tel que présenté par Tutin et Grossmann (2014 : 19), les linguistes qui ont piloté cette entreprise au sein du laboratoire LIDILEM de l'Université de Grenoble 3-Stendhal.

CONCLUSION

Pour résumer, l'analyse linguistique du DS se base sur des productions langagières en situation d'usage, c'est-à-dire sur des données qui ont été produites dans des situations particulières. De fait, le recueil de ces données impose une contextualisation prenant en compte une panoplie de paramètres, comme les types de locuteurs, leur statut dans la communauté discursive, la culture dont ils font partie, la (sous)discipline, les genres de textes, la datation, etc. Ces paramètres d'ordre extralinguistique sont cruciaux dans l'interprétation, c'est-à-dire dans la construction du sens. Autrement dit, plus l'analyste veille à ce que ces paramètres soient respectés, plus les résultats de son analyse ont de la chance d'être généralisables. *Scientext*, ce vaste corpus d'écrits scientifiques, peut, nous semble-t-il, être utile dans ce sens. Il peut, en effet, offrir la possibilité au chercheur de prendre en compte ces paramètres et garantir en quelque sorte la scientificité de l'étude menée et des résultats obtenus. Mais la question qui se pose est jusqu'à quel point on peut neutraliser ces variables, notamment au vu des marges que certains auteurs se permettent par rapport aux normes régissant le même genre, la même (sous)discipline, la même culture, etc.

---

[1] Cité dans Vaguer (2007 : 222)

[2] Dans Corbin (1980), la première façon de faire relève de la « linguistique de terrain », tandis que la seconde, étant introspective, relève de la « linguistique de bureau ». Dans le domaine du traitement automatique des langues (TAL), cette dichotomie renvoie respectivement aux « TAL robuste » et « TAL théorique » (voir Cori, 2008).

[3] Voir Vaguer (2007) pour une description détaillée des avantages et inconvénients de ces deux démarches.

[4] Nous entendons par là les phrases complexes, par opposition aux phrases simples.

[5] Le contexte dans son sens englobant, c'est-à-dire le texte, le cotexte et le contexte extra-linguistique.

[6] L'analyse quantitative s'effectue généralement sur un corpus de grande taille.

[7] Habilitation à diriger des recherches.

[8] Les sciences dites « dures » sont les sciences exactes, par opposition aux sciences de l'homme, qui sont appelées aussi « sciences douces », ou encore « molles » dans un sens péjoratif. Pour Grossmann (2012 : 143-144), ces deux familles renvoient respectivement à deux modèles de scientificité : les *modèles de prédiction* et les *modèles herméneutiques*.

[9] L'homogénéité du genre de l'article est également discutable. Swales (2004) le scinde en trois sous-genres distincts : l'article expérimental, l'article théorique et l'article de synthèse.

[10] « La science a été longtemps perçue comme le reflet de la vérité, et la langue, medium profane utilisé par un auteur faillible, comme un obstacle à l'expression de vérités scientifiques » (Poudat, 2006 : 49)

[11] Au sens de Swales (1990), en l'occurrence la communauté des chercheurs.

[12] Les fonctions auctoriales indiquées *supra* ne sont pas exclusives.
[13] Librement interrogeable sur https://corpora.aiakide.net/scientext20/?do=SQ.setView&view=corpora

RÉFÉRENCES BIBLIOGRAPHIQUES

ADAM Jean-Michel, (1997), « Genres, textes, discours : pour une reconception linguistique du concept de genre», *Revue Belge de Philologie et d'Histoire*, n°75.

BOCH Françoise et GROSSMANN Francis, (2002), « Se référer au discours d'autrui, quelques éléments de comparaison entre experts et néophytes », *Enjeux*, n°54.

BOURE Robert, (1998), « Produire une revue scientifique. Le cas de Sciences de la Société » *in* Renzetti, Françoise (coord.), *Stratégies informelles et valorisation de la recherche scientifique publique*, ADBS, 1998, pp. 105-119.

CHARAUDEAU Patrick, (2009), « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus*, n°8, URL : http://corpus.revues.org/1674 (consulté le 13 mars 2023).

CHARAUDEAU Patrick et MAINGUENEAU Dominique, (2002), *Dictionnaire d'analyse de discours*, Le Seuil, Paris.

CORBIN Pierre, (1980), « De la production des données en linguistique introspective », *Théories linguistiques et traditions grammaticales*, PUL, Lille.

CORI Marcel, (2008), « Des méthodes de traitement automatique aux linguistiques fondées sur les corpus », *Langages*, n° 171.

DALBERA Jean-Philippe, (2002), « Le corpus entre données, analyse et théorie », *Corpus*, n°1, URL : https://journals.openedition.org/corpus/10#:~:text=38Bref%2C%20le%20corpus%20du,selon%20divers%20modes%20et%20rassembl%C3%A9es (consulté le 13 mars 2023).

DUBOIS Jean et SUMPF Joseph, (1969), « Problèmes de l'analyse du discours », *Langages*, n°13.

FALAISE Achille , GROSSMANN Francis, KRAIF Olivier et TUTIN Agnès, (2009), « Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques », *Journées internationales de linguistique de corpus*, Lorient.

FLØTTUM Kjersti, (2008), *Language and discipline perspectives on academic discourse*, CSP, Cambridge.

FLØTTUM Kjersti et VOLD Eva Thue, (2010), « L'éthos auto-attribué d'auteurs-doctorants dans le discours scientifique », *Lidil*, n°41.

GEISLER Cheryl et KAUFER David, (1989), « Novelty in academic writing », *Written communication*, n°6.

GLEASON Henri-Allan, (1969), *Introduction à la linguistique*, Larousse, Paris.

GROSSMANN Francis, (2012), « Pourquoi et comment cela change ? Standardisation et variation dans le champ des discours scientifiques », *Pratiques*, n°153/154.

GROSSMANN Francis et TUTIN Agnès, (2014), *L'écrit scientifique : du lexique au discours*, Presses Universitaires de Rennes, Rennes.

MAYAFFRE Damon, (2005), « Les corpus politiques : objet, méthode et contenu. Introduction. », *Corpus*, n°4, URL : https://journals.openedition.org/corpus/292 (consulté le 13 mars 2023).

MELLET Sylvie, (2002), « Corpus et recherches linguistiques. Introduction », *Corpus*, n°1, URL : http://corpus.revues.org/7 (consulté le 13 mars 2023).

NORMAND Ariane, (2014), « Proposition pour l'induction en analyse du discours », *Approches inductives*, n°1.

POUDAT Céline, (2006), *Étude contrastive de l'article scientifique de revue linguistique*, Thèse de doctorat, Université d'Orléans.

REUTNER Ursula, (2010), « *De nobis ipsis silemus ?* Les marques de personne dans l'article scientifique », *Lidil*, n°41.
RINCK Fanny, (2006), *L'article de recherche en Sciences du Langage et en Lettres, Figure de l'auteur et approche disciplinaire du genre*, Thèse de doctorat, Université Grenoble 3.
RINCK Fanny, (2010), «L'analyse linguistique des enjeux de connaissance dans le discours scientifique. Un état des lieux », *Revue d'Anthropologie des connaissances*, n°3.
ROUVEYROL Laurent, (2005), « Vers une logométrie intégrative des corpus politiques médiatisés. L'exemple de la subjectivité dans les débats-panel britanniques », *Corpus*, n°4, URL : http://corpus.revues.org/293 (consulté le 14 mars 2023)
SWALES John, (1990), *Genre Analysis : English in Academic and research settings*, Cambridge University Press, Cambridge.
SWALES John, (2004), *Research Genres. Exploration and Applications*, CUP, Cambridge.
TUTIN Agnès, (2010), « *Dans cet article, nous souhaitons montrer que …* Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines. Enonciation et rhétorique dans l'écrit scientifique », *Lidil*, n°41.
VAGUER Céline, (2007), « Corpus, vous avez dit corpus : De la notion de corpus à la création d'un corpus informatisé », *Corpus, Langues et Linguistique*, Actes des *3e Journées de la linguistique de corpus*, URL : https://www.cairn.info/load_pdf.php?ID_ARTICLE=LF_211_0011&download=1 (consulté le 14 mars 2023).

ANNEXES

Tableau comportant les principales informations sur *Scientext*, en particulier les écrits scientifiques en français.

Discipline	**Genre d'écrit**	**Volume**	**Annotations**	**Disponibilité**
Pour le corpus public : linguistique, psychologie, sciences de l'éducation, TAL, biologie, médecine, électronique, mécanique.	Articles de recherche, communications écrites, thèses de doctorat, mémoires d'HDR.	4,8 millions de mots pour le corpus public (en ligne). 8,4 millions de mots pour le corpus interne de l'équipe.	Syntaxiques et morphosyntaxiques. Annotation structurale détaillée des parties textuelles.	Corpus public interrogeable en ligne. Corpus interne accessible en intranet.

POUR CITER L'AUTEUR :

NAJAI Noureddine, (2024), « Autour du corpus exploité dans l'étude du discours scientifique », Ex Professo, V09, N01, pp. 08- 19, Url: https://www.asjp.cerist.dz/en/PresentationRevue/484