Les corpus textuels numériques (re)spécifiés

Damon Mayaffre

Les corpus textuels numériques (re)spécifiés

Damon Mayaffre, Université Côte d'Azur, CNRS, BCL-UMR 7320, France - damon.mayaffre@univ-cotedazur.fr

Introduction

Les années 2000 ont vu le triomphe du corpus en linguistique et, par-delà, dans les SHS. Non pas que l'objet corpus n'ait existé de longue date auparavant, non pas que les linguistes l'aient ignoré jusqu'alors mais au sens où la linguistique *sans* ou *hors corpus* apparaît à ce moment-là comme une spéculation intellectuelle marginale pratiquée seulement par une minorité. Sémanticiens, phonologues, lexicologues, dialectologues, etc. se revendiquent tous du corpus ; même la syntaxe générative semble concernée comme par exemple dans le numéro « La syntaxe de corpus » de la revue éponyme *Corpus* au début du siècle.

Et la fièvre *corpus* qui a saisi l'hexagone scientifique, sinon le monde, au début des années 2000 n'est jamais retombée, attestant qu'il s'agissait plus qu'un effet de mode. On lira ainsi à titre d'exemple un premier bilan documenté dans [Laks 2008], on consultera les actes du colloque thématique du Cercle belge de Linguistique [Mellet et Longrée 2009] ; on mentionnera le 23ème colloque international du Cercle Linguistique du Centre et de l'Ouest (Université de Poitiers - 5 et 6 juin 2009) intitulé « L'exemple et le corpus : quel statut ? ». A l'échelle internationale le *peer-reviewed journal* intitulé *Corpora*, créé en 2006, fait paraitre son 15ème volume en 2020. A l'échelle nationale, la revue *Corpus* créé en 2001 fait paraitre ses 20ème et 21ème numéros cette même année. A Lorient, Poitiers, Grenoble, les Journées annuelles de Linguistique de Corpus (JLC) ont tenu leur 10ème session en 2019, et les Journées internationales biannuelles d'Analyse de Données Textuelles (JADT) se sont imposées dans le concert national et européen avec un millésime à Nice en 2016, à Rome en 2018 et à Toulouse en 2020. Et ces indices sur l'engouement *corpus* ne sont pas exhaustifs, puisque nous pourrions ajouter, autres exemples, la parution récente de *La Mesure et le Grain. Sémantique de corpus* de François Rastier (2011) ou de *Explorer un corpus textuel : Méthodes - pratiques - outils* de Frédéric Landragin et Céline Poudat (2017), prolongement certain, 20 ans après, de l'ouvrage de Benoît Habert, Adeline Nazarenko, André Salem, *Les linguistiques de corpus* (1997).

Cependant, loin de tirer un bénéfice direct du triomphe du corpus, la linguistique de corpus *stricto sensu* au sens par exemple de [Aijmer and Altenberg (ed.) 2002 ; Biber, Conrad & Reppen 1998 ; Habert, Nazarenko et Salem 1997 ; Kennedy 1998, Partington 1998 et 2004, Sinclair 1991, Tognini-Bonelli 2001 ; Rastier 2001, 2005-a, 2005-b, 2011, Williams 2005, Poudat et Landragin 2017 ; etc.][1] s'en trouve ébranlée comme si, par une ruse de l'histoire scientifique, la banalisation de son objet avait brouillé voire dissout son identité propre. Pire : revendiqués désormais par tous, les corpus semblent ne plus appartenir scientifiquement à personne jusqu'à devenir inopérants. *A minima* l'objet *corpus* demande aujourd'hui à être (re)spécifié.

[1] Il va de soi que ces auteurs divergent entre eux sur certains points. Nous les rassemblons ici en tant que linguistes ayant pris à bras le corps l'objet *corpus* et l'idée d'une linguistique de corpus. Quelques grands ancêtres pourraient être ajoutés comme Firth ou Palmer.

C’est donc dans ce cadre paradoxal – triomphe des corpus ; dilution de la linguistique de corpus – que cette contribution essaye de (re)questionner l’objet *corpus textuel numérique*, dans sa dimension textuelle comme dans sa dimension numérique. Les 5 portraits proposés (1. *Le corpus comme matrice* ; 2. *Le corpus comme contexte* ; 3. *Corpus réflexifs et herméneutique endogène* ; 4. *Le corpus comme texte* ; 5. *Corpus et numérique*) tendent vers une exigence méthodologique forte qui sera partout suggérée sans pouvoir être hélas aboutie nulle part dans cette contribution. Dès lors qu’il est théoriquement établi et empiriquement constitué, tel que nous allons essayer de le discuter, comment traiter scientifiquement un corpus textuel numérique ? Quelles méthodes et quels logiciels utiliser ? Le *deep learning* aujourd’hui peut-il compléter les parcours de la logométrie traditionnelle ? etc. Ces aspects méthodologiques introduits dans cet article sont traités ailleurs, et nous ne pouvons renvoyer ici le lecteur qu’à la bibliographie (pour commencer, peut-être, trois ouvrages récents et un article : Lebart *et al*. 2019, Née *et al*. 2018, Mayaffre *et al*. 2019, Mayaffre *et al*. (sous la dir.) 2020b).

1.1. Le corpus comme matrice

Travailler sur corpus, c’est vouloir travailler sur des données attestées ; sans quoi l’introspection suffirait. Comme les seules données langagières attestées – c'est-à-dire les seules performances linguistiques abouties qu’un locuteur produit lorsqu’il s’exprime – sont les discours[2] puis, en tant que formes empiriques et stabilisées du discours, les textes, nous ne considérons, dans cette contribution, que les corpus textuels[3].

Et peut-être est-ce la nature nécessairement textuelle des corpus traités qui modifie en substance les choses et notre réflexion ?

Avec François Rastier, nous pensons qu’un texte n’a pas de signification, qu’une grammaire formelle du texte est vaine, que les unités textuelles (quand bien même reposent-elles sur les formes matérielles graphiques repérables) sont mouvantes, plurielles, complexes, parfois discontinues et opèrent à différents niveaux de granularité linguistique (lexique, grammaire, syntaxe, graphie, pragmatique). Un texte n’a pas de signification mais un sens (ou plutôt des sens) qu’il ne s’agit pas de re-trouver mais de co-construire dans des parcours de lecture contrôlés. En linguistique du texte, il s’agit donc moins d’établir ou de restituer, que d’interpréter[4].

Partant, un corpus de textes n’est pas une *base de données* à interroger : le sens ne se laisse pas enfermer dans une *base ;* le sens n’est jamais *donné*. Les corpus textuels apparaissent ainsi très différents de certains corpus-ressources lexicographiques ou phonologiques qui consignent dans de vastes tableurs leurs données. Par la nature de ses composants – les textes

[2] « Discours » au sens large donc, c’est-à-dire aussi bien des monologues que des dialogues ou des séquences interactionnelles longues ou courtes.

[3] Répétons : un locuteur ne dit pas des mots isolés, il ne fabrique pas des phrases grammaticales : il produit des discours, aussi courts soient-ils, qui prennent la forme empirique de textes analysables.

[4] La dimension herméneutique de la linguistique (textuelle) est, on le comprend dès à présent, au cœur de cette contribution. Elle est au centre de l’œuvre de Rastier et particulièrement de [Rastier 2001]. Nous sommes marqué par le programme qui ouvre le chapitre IV *Herméneutique matérielle* : « Il reste à unir, au sein d’une sémantique des textes, les acquis de la philologie et de la linguistique comparée, pour restituer aux sciences du langage leur statut de disciplines herméneutiques » (p. 99). Nous retrouvons le même type de programme sous la plume de [Adam 2008 : 30] lorsque l’auteur parle « du tournant herméneutique et plus largement de l’ouverture de la linguistique à l’interprétation ». Ajoutons que notre dette scientifique a été contractée, en amont de ces linguistes du texte, envers Jacques Guilhaumou [Guilhaumou, 2006]. En France, c’est lui qui a engagé précocement l’école française d’Analyse du discours dans ce *tournant herméneutique* décisif.

–, le corpus textuel n'est donc pas une banque[5] ou une base de *data*. C'est un lieu, lui-même construit, où s'échafaude le sens, où se scénarise l'interprétation. En d'autres termes, nous pensons que le corpus est *moins le réceptacle du sens que sa matrice* ; moins un observatoire d'une langue qui serait déjà-là, qu'un observé vivant, mouvant, dynamique qui par sa constitution même et par son organisation, produit un sens toujours à inventer.

Cette première affirmation semble avoir des conséquences majeures. Nous en relèverons ici succinctement deux seulement.

D'un point de vue méthodologique d'abord, la méthode pour traiter ce corpus-matrice – la logométrie ou le deep learning notamment – prendra une valeur heuristique plus que probatoire : interroger plutôt que prouver, interpréter autant qu'établir. Notre travail s'inscrit à l'intérieur d'une linguistique de corpus à vocation herméneutique – l'herméneutique matérielle numérique[6] – et non dans le traitement automatique de la langue. Si la logométrie est, loin de l'impressionnisme, une méthode formalisante s'appuyant fermement sur le matériel du texte, elle formalise moins des données que des parcours interprétatifs : la nuance est fondamentale.

D'un point de vue épistémologique ensuite, le corpus-matrice redéfinit notre posture face au texte, et le sens de la démarche. A moins d'aspirer à une démarche paradoxale et circulaire, qui consisterait à supposer (à « hypothéser ») un sens déjà-là qu'il suffirait de rechercher et d'établir, tout en prétendant que le corpus produit un sens qu'il nous resterait à inventer et à co-construire, notre mode heuristique demande à être renversé. *A minima*, nous parlerons, avec [Tognini-Bonelli 2001] et avec toute la littérature anglo-saxonne, d'études *corpus-driven* (*versus* d'études *corpus-based*) : un corpus qui n'est pas une ressource que l'on soumet à l'interrogatoire mais un objet qui nous dirige dans le questionnement. Plus hardiment, nous parlerons d'un retournement de la méthode hypothético-déductive qui domine en SHS : là où l'on interrogeait *top-down* le corpus, l'on se propose en effet de se laisser interroger *buttom up* par lui. Posons ici le principe naïvement en renvoyant à [Mayaffre 2010, chapitre 2] pour le détail.

1.2. Le corpus comme contexte

Matrice du sens, le corpus l'est car, dans son entier, il informe les textes qui le composent. Principe d'architextualité ou *détermination du local* (dans ce cas, le texte) *par le global* (dans ce cas, le corpus) dirait [Rastier 2001 : 92] ; condition d'une comparaison différentielle non hiérarchisée des textes diraient [Adam et Heidmann 2005 : 102 et ss.].

Soulignons simplement que cette affirmation théorique et récente, que la plupart des auteurs de linguistique textuelle ou de linguistique de corpus revendiquent désormais, épouse le principe même, technique et originel, de l'analyse de données textuelles, de la lexicométrie ou de la logométrie dès les années 1960 : l'idée d'un corpus-norme ou d'un corpus-référence, l'idée d'une statistique endogène ; comme si cette méthode était idéale pour ces linguistiques, comme si la méthode avait devancé, en pareil cas, la théorie. A la statistique endogène de

[5] Ainsi, contrairement à la terminologie anglo-saxonne, nous distinguons clairement *banque* de textes (ressource matérielle collective dans laquelle le chercheur pourra puiser ses textes) et *corpus* de textes (objet construit de manière *ad hoc* par lequel le chercheur problématise sa recherche). Le *British National Corpus*, le *Brown Corpus* ou le *Lancaster-Oslo-Bergen (LOB) Corpus* relèvent pour nous des banques de textes au même titre que Frantext, la BMF ou le LASLA.

[6] Voir le titre de notre thèse HDR dont cet article est largement issu : D. Mayaffre, *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, 3 vol. 107, 232, 414 p. Soutenue à Nice, le 30 avril 2010.

[Guiraud 1954], [Muller 1977] ou [Brunet 2011] a en effet répondu une « stylistique endogène » [Viprey 1997], ou une « lexicologie textuelle » endogène au texte [Valette 2008] : finalement, la linguistique de corpus n'est rien d'autre qu'une *linguistique endogène*, et son outillage par la statistique (endogène par essence) apparaît naturel aussi bien chez [Biber 1988, 1995, 1998], [Habert, Nazarenko et Salem 1997] que [Malrieu et Rastier 2001].

Matrice du sens, le corpus entretient, dès lors, un dialogue direct avec la notion de *co(n)texte* puisque l'on admet que la co(n)textualisation est la condition de la maïeutique du sens. Le sens naît en/du co(n)texte, avons-nous plusieurs fois écrit [Mayaffre 2007b et 2014] en soulignant la vanité d'une linguistique ou simplement de pratiques méthodologiques décontextualisantes. Le sens naît en/du corpus pourrait-on renchérir ici liant corpus et contexte dans une relation étroite, quasi-synonymique.

Le corpus peut être en effet conçu comme une forme privilégiée du co(n)texte. Plus précisément, nous définissons le corpus comme *la forme maximale du co(n)texte*. De la lettre au mot, du mot à la phrase, de la phrase au paragraphe ou à la partie, de la partie au texte, du texte au corpus : le phénomène de co(n)textualisation (ou, autrement dit, l'extension de l'objet du linguiste) semble devoir aller jusque là et s'arrêter là.

Précisons bien cependant pour ne pas paraître naïf : en définissant le corpus comme forme maximale du co(n)texte, nous entendons forme maximale *formalisable* du co(n)texte, car le co(n)texte (le co-texte proche ou l'intertexte plus lointain) est insondable et à proprement parler *insaisissable*. Plus loin encore, le contexte lorsqu'il s'étend au-delà du co-texte ou de l'intertexte pour toucher à la situation socio-historique générale et aux conditions de production des discours est une chose qui échappe pour partie aux études strictement linguistiques ; le *hors corpus* nous mène vers un horizon scientifique insondable.

Pour un texte donné, donc, une infinité de corpus pourront être construits formant autant de co(n)textes maximaux formels, au sein desquels seront écrits autant de scénarios interprétatifs. Le corpus donne un corps – un corps linguistique – au contexte, et hors de la réification qu'il propose, le co-texte ou le contexte, l'intertextualité ou l'interdiscursivité sont des magmas sans forme, sans aucun doute passionnants mais platoniques ou insaisissables. Dit plus simplement, le corpus représente pour nous la forme empirique ou matérielle du contexte ; celle qui accepte de se soumettre à l'observation linguistique. C'est le cadre matériel, immédiat, formalisé, d'une interprétation contrôlée.

1.3. Corpus réflexifs et herméneutique endogène

Pour répondre à cette double définition – le corpus comme matrice du sens ; le corpus comme objectivation du contexte linguistique nécessaire à l'interprétation–, les corpus textuels doivent être bien formés (Landragin et Poudat 2017).

Outre les critères désormais bien connus d'équilibre, de représentativité ou d'exhaustivité, d'homogénéité (notamment générique), de contrastivité et de clôture, etc., nous avons proposé non seulement que les corpus soient gros pour offrir un cadre suffisant[7] mais structurés de manière adéquate.

[7]A vrai dire, la taille des corpus est une question insoluble, inutilement polémique. Nous savons que le traitement statistique est d'autant moins contestable que les populations sont importantes, et la puissance des machines repousse chaque jour les limites de la veille. Aussi « more data, better data » pourrait être une devise pertinente s'il n'y avait la nécessité d'embrasser le texte qualitativement : trop gros, les corpus deviennent illisibles et ininterprétables. Puisqu'il faut donner un ordre de grandeur, posons, pour ce qui nous concerne, que nous

La proposition qui a été la mieux reprise par la communauté scientifique est la nécessaire *réflexivité du corpus*. Sans revenir dans le détail sur une idée développée par le menu dans (Mayaffre 2002) et (Mayaffre 2007b) et déclinée dans (Mayaffre et al. 2020a et Mayaffre 2020c), posons dans la continuité de ce qui précède que l'enjeu du *corpus réflexif* est de constituer un ensemble sémantique auto-suffisant qui internaliserait les ressources co-textuelles nécessaires à l'interprétation de chaque texte. En miroir, les textes du corpus doivent s'éclairer mutuellement ; se *réfléchir* les uns les autres ; chacun d'entre eux constituant le co-texte immédiat de tous, et l'ensemble constituant l'intertexte de chacun.

Comme l'importance de la notion a déjà été soulignée précédemment, avançons ici l'idée que le *corpus réflexif* est la condition d'une herméneutique *endogène*. C'est au sein du corpus que les parcours interprétatifs sont proposés ; le corpus réflexif dans son ensemble formalisant l'intertextualité – une intertextualité parmi d'autres possibles – des textes constituants, soumis à l'analyse.

Internaliser les ressources interprétatives dans des corpus réflexifs, pour une herméneutique endogène : l'ambition paraîtra démesurée mais rappelons le modeste point de départ – un malaise épistémologique – de la réflexion. Une discrimination non acceptable sépare souvent les textes-objets-du-corpus et les textes-ressources-interprétatives-hors-corpus. Les premiers font l'objet d'un mode de sélection, d'une attention philologique et, par définition, d'un traitement linguistique minutieux. Les seconds sont convoqués à discrétion, cités à la hussarde et, hors du corpus, échappent au traitement proprement dit. Cette discrimination saute à l'œil car, sources ou ressources, il s'agit bien dans les deux cas de *textes.*

Notons que ce co-texte ou cet intertexte que les corpus réflexifs entendent manufacturer peuvent apparaître à l'usage objectif : il s'agit de traiter ensemble deux textes contemporains jugés comme apparentés historiquement – le premier constituant le co-texte immédiat objectif du second, le second le co-texte immédiat objectif du premier –, comme par exemple, dans nos travaux, les discours de Jospin chef de gouvernement et les discours de Chirac président de la République, durant la période de cohabitation (ie. explicitement les deux locuteurs se répondent). Mais la réfléxivité mise en scène peut être plus subjective et arbitraire, et mettre volontairement en dialogue deux textes « étrangers » l'un à l'autre, mais dont on suppose, par hypothèse, que la mise en rapport – le face à face *réflexif* – au sein du corpus produira des effets de sens et suscitera des parcours de lecture critiques et fertiles. Précisons par-là, avec force, combien le corpus est un objet construit sur la base d'hypothèses de travail, et combien cette construction détermine l'analyse. Avec des méthodes *corpus-driven* ou émergentistes comme la logométrie ou le deep learning, les hypothèses de travail ne doivent pas présider au traitement linguistique du corpus (laissons remonter librement et sans a priori du corpus des informations linguistiques pertinentes) : il est suffisant et impérieux que les hypothèses de travail président au recueil – *moment philologique* [Adam et al. 2005 : 83] – et à l'organisation réflexive – *projection herméneutique* – des textes constituant du corpus.

1.4. Le corpus comme texte

De la lettre au corpus, en passant par le mot, la phrase ou le paragraphe, et en s'arrêtant sur le texte, la linguistique de corpus procède à/de l'extension de l'objet de la linguistique vers des réalités ou des globalités toujours plus vastes et toujours plus complexes ; les corpus représentent pour nous le *terminus ad quem* d'une linguistique contextualisante.

---

traitons plutôt des corpus de 2.000.000 de mots que de 200.000 (trop petits pour mesurer des régularités) ou de 20.000.000 (trop gros pour être lisibles).

Par facilité sans doute, à chaque palier de complexité franchi, l'analyste a eu tendance à se retourner vers le palier inférieur pour en faire remonter des schémas d'analyse qui lui étaient familiers. Ainsi, hier, par exemple, lorsqu'on est passé du phrastique au transphrastique, s'est-on imaginé établir – en vain – une *grammaire du texte* comme il en existait une précédemment de la phrase.

Ainsi, aujourd'hui, en passant du texte au corpus peut-on envisager expliquer – avec fruit ? – la « corporalité » comme on explique la textualité ; et précisons que la tâche s'annonce passionnante mais d'autant plus compliquée que la notion de textualité est elle-même à peine stabilisée en linguistique textuelle et objet encore de riches discussions.

Peut-on considérer un corpus comme un texte ? Peut-on considérer un corpus textuel comme un macro-texte qu'il s'agirait alors de traiter avec des outils théoriques en partie balisés par Hjelmslev ou Bakhtine, Hasan, Halliday, Adam ou Rastier ?

Nos travaux se gardent bien d'apporter une réponse définitive à une interrogation qui engage sinon l'avenir de la linguistique de corpus en tout cas son intérêt actuel. Mais plusieurs indices, comme autant de pistes de réflexion, peuvent être pressentis. Deux méritent d'être rappelés ; nous verrons qu'ils sont nuancés ; et que dans ces nuances réside des programmes de recherche.

Cohérence-cohésion du texte / cohérence-cohésion du corpus

Si l'on peut supposer que le corpus, à l'image du texte, est un ensemble cohérent et cohésif, il l'est nécessairement de manière différente. En s'aventurant dans le distinguo heideggerien sans doute peut-on prétendre que la cohésion-cohérence d'un texte lui est ontologique ; la cohésion-cohérence du corpus lui est ontique. Le texte est cohérent par nature, par essence, par définition[8] ; le corpus, lui, l'est par existence – en tant qu'*étant*–, par construction, par hypothèse.

Il ne s'agit pas ici de naturaliser (ontologiser) l'objet texte qui est lui-même un objet construit, artefactuel, mais de rappeler que le texte existe – sous une forme ou une autre – dans la société sans l'analyse scientifique, là où le corpus existe uniquement en laboratoire par le fait du seul chercheur, et seulement le temps de la recherche[9]. Le texte est un construit, mais un construit social ou culturel « de première main » par le fait du couple auteur-lecteur. La construction du corpus textuel est, elle, de « seconde main » par le seul fait de l'analyste.

La textualité – ce qui fait qu'un texte est un texte – est définitoire du texte et peut être perçue par tout lecteur, *sans quoi il n'admettrait pas qu'il s'agit là d'un texte*. La corporalité, elle, est une pétition de principe ou un parti pris, un espoir, un postulat, le fruit d'un travail singulier ou d'une projection particulière évidente pour le seul chercheur. Exprimée en langage mathématique, la cohésion-cohérence d'un texte est axiomatique ; la cohésion-cohérence du corpus est hypothétique. Bref, un texte qui ne serait pas cohérent-cohésif ne serait plus un

[8] Les « propos incohérents » qu'essayent de tenir certains auteurs n'en peuvent mais. C'est précisément cette incohérence étudiée du propos qui fait la cohérence du texte.

[9] Rappelons les échecs répétés d'archives de corpus ou de banques de corpus. Tous les chercheurs ont rêvé de sauvegarder et patrimonialiser leurs corpus (et non seulement les textes) mais force est de constater que ces tentatives sont le plus souvent vaines. Les corpus semblent destinés à disparaître après l'analyse et la validation de l'étude par des jurys. Certes, des *benchmark corpus* existent pour comparer des méthodes et hiérarchiser des logiciels mais il ne s'agit pas de corpus SHS en vue d'une analyse.

texte. Un corpus qui n'est pas cohérent-cohésif est seulement un corpus manqué, c'est-à-dire manquant de pertinence et d'efficacité heuristique ; mais cela reste un corpus.[10]

La différence est donc importante. Pourtant, elle n'est pas définitive. Quoique d'une autre nature, la cohérence-cohésion du corpus est l'enjeu de la linguistique de corpus exactement comme la cohérence-cohésion du texte est celui de la linguistique textuelle : c'est cette tension commune vers une textualité / corporalité, conçue avec [Charolles 1995 : 10] comme « **principe général** gouvernant l'interprétation »[11] du texte / corpus, qui rapproche les deux disciplines. Simplement, si du point de vue de la cohérence-cohésion du texte, de [Halliday et Hasan 1976] à [Calas 2006] ou [Adam 2008], l'essentiel est déjà réalisé, du point de vue du corpus, l'essentiel reste à faire, même si les travaux de [Viprey 1997, 2005, 2006] sur la micro/macro-distribution des unités dans le corpus et la *texture* des corpus balisent une partie du terrain. Et à ce stade, pressentons seulement, d'un point de vue méthodologique, que le parallèle texte / corpus et textualité / corporalité demande quelques ajustements : si le texte et la textualité peuvent encore être considérés comme des objets *micro* réclamant l'approche qualitative, le corpus et la corporalité, en tant qu'objets *macro*, semblent exiger une approche quantitative.

Sérialité du corpus / linéarité du texte

Si texte et corpus (textuel) présentent certaines similarités au point que l'on peut envisager entre eux un simple rapport d'échelle, une différence profonde de structure semble les distinguer : le corpus est fondamentalement un objet *sériel* (Mayaffre 2002), le texte est d'abord un objet *linéaire*.

Le corpus est une *collection* de textes réunis sur la base d'hypothèses de travail. Au-delà du stade critique d'une collection de textes qui en compterait un seul, les corpus peuvent donc être considérés comme des séries. (Et faut-il souligner encore ici que les séries, en linguistique comme ailleurs, se prêtent bien au traitement statistique ?)

Certes, certaines de nos séries, particulièrement en histoire, sont ordonnées linéairement. Nous pensons aux *séries textuelles chronologiques* dont André Salem a décrit les caractéristiques [Habert, Nazarenko, Salem 1997 : 207 et ss] et qui, précisément, par leur *progression* chronologique, et la permanence de leurs locuteurs individuels ou collectifs, peuvent à juste droit être traitées comme des textes : il s'agit-là d'un champ de recherche à part entière de la linguistique de corpus dont [Viprey 2004] ou [Metwally 2017], sur les numéros du *Monde diplomatique*, échelonnés sur plusieurs décennies, ont décrit le fonctionnement.

Mais hors des séries textuelles chronologiques, la plupart des corpus n'ont pas de structure linéaire évidente ; de manière significative leurs parties (les textes qui les composent ou des regroupements de textes que l'on aura constitués en parties) peuvent être indifféremment ordonnées sans que le traitement en soit changé. Ainsi dans le corpus de la campagne électorale de 2007 ou de 2017 que nous avons eu l'occasion de traiter, les textes des candidats Laguiller, Buffet, Royal, Bayrou, Sarkozy et Le Pen ou Mélenchon, Hamon, Macron, Fillon,

[10] Le lecteur aura remarqué le parti pris de mentionner ensemble, globalement, la *cohérence* et la *cohésion*. Dans le détail, et de manière hiérarchique, il serait facile de montrer que la *cohésion* du corpus pose plus de problème encore que sa *cohérence*. Cf. infra la question de la sérialité des corpus (*versus* la continuité des textes).

[11] Surligné par Michel Charolles. Voir aussi son article moins abouti de 1983 : Coherence as a principle in the interpretation of discourse. *Text*, 3-1, 71-99.

Le Pen peuvent contraster et se singulariser indépendamment de leur ordre de saisie [Mayaffre et al. 2017 et Mayaffre 2020c].[12]

Si le corpus est avant tout sériel donc, et non nécessairement organisé linéairement, le texte lui est toujours linéaire ; c'est un objet linéaire. A l'exception de quelques productions surréalistes ou de quelques jeux d'auteurs marginaux en littérature[13], un texte a toujours un commencement, un prolongement et une fin ; il peut être défini comme une *suite* [Maingueneau 1996 : 81 ou Détrie, Siblot, Vérine 2001 : 349]; et l'élément fondamental de sa lecture est sa progression, pour nous de gauche à droite, de haut en bas. Contrairement aux parties du corpus, les parties du texte (ses phrases, ses chapitres, ses séquences…) ne peuvent être inversées sans remettre en cause l'édifice. Certes, depuis l'abandon du rouleau pour le codex ou le *polyptychon*, rien n'interdit au lecteur de briser par sa lecture cette progression implacable et de papillonner aléatoirement d'une page à l'autre, d'arrière en avant, ou de chapitre en chapitre. Certes encore, aujourd'hui, le numérique permet des lectures hypertextuelles dont la caractéristique est justement de s'affranchir du linéaire : il s'agit d'une révolution majeure qui est au cœur même des propositions méthodologiques du traitement informatique et statistique des textes et sur lesquelles nous reviendrons. Mais il n'en reste pas moins vrai que la linéarité apparaît irréductible à la textualité et constitue le socle de sa définition[14].

Suite continue *versus* série discontinue, linéarité du texte *versus* sérialité du corpus : touchons-nous donc cette fois-ci à une différence définitive ? Non pourtant.

Objet linéaire, *d'abord*, le texte est aussi traversé de sérialité et de réticularité : c'est l'apport essentiel des travaux de [Viprey 1997, 2005, 2006] que nous avons essayé de reprendre d'un point de vue théorique et pratique dans l'ensemble de nos écrits [pour une approche globale : Mayaffre 2010]. Objet sériel, *en premier*, le corpus est – ne serait-ce que par ce qu'il est composé de textes linéaires – traversé par la linéarité et la séquentialité : c'est l'apport essentiel par exemple des travaux de [Mellet et Longrée 2009] ou [Longrée et Mellet 2013]

Autrement dit, se dessine un double mouvement scientifique qui venant de deux pôles opposés converge en un programme de recherche commun : la linguistique de corpus, partant de la série, vise aujourd'hui à réintroduire la linéarité dans ses traitements. La linguistique textuelle, partant du linéaire, admet la sérialité comme élément complémentaire de son objet. A la suite de Jean-Michel Adam qui, venant du texte, a présenté l'unité de ce programme aux JADT 2006 [Adam 2006] avant d'en esquisser des pistes dans [Adam 2008 : 179-182], nous avons essayé, venant du corpus, d'en souligner quelques enjeux [Mayaffre 2007a] ou [Mayaffre 2014].

Dit en des termes admis depuis longtemps en linguistique, il s'agit de rappeler que toute écriture / lecture (celle d'un texte ou celle d'un corpus ; *a fortiori* celle d'un *corpus textuel*) articule une compétence syntagmatique et une compétence paradigmatique. Longtemps essentiellement paradigmatique, l'approche statistique des données textuelles doit prendre en

---

[12] Cette idée pourrait être nuancée (mais non contredite). On aura en effet noté qu'un *ordre* politique ici s'est imposé à nous. Et lors de l'analyse, certaines distributions semblent renvoyer à cet ordre ou cette *progression* du corpus. Ainsi par exemple constatera-t-on une progression de l'emploi de « patrie » à mesure que le corpus se *déroule* (au sens politique *et* typographique) vers la droite.

[13] Précisément il s'agit là de jeux, dont la règle sous-jacente est bien la linéarité attendue... et transgressée.

[14] Citons ici la concession définitive du linguiste qui a remis le mieux en cause cette linéarité pour introduire dans le traitement la réticularité : « Nul ne saurait mettre en doute qu'un texte se manifeste dans l'ordre du temps et/ou de l'espace orientés, se caractérise par un début, un milieu, une fin, ordonnés et non interchangeables, et ce à quelque échelle que ce soit de l'organisation macro-séquentielle à la fine succession des périodes » [Viprey 2006 : 74].

compte désormais aussi la dimension syntagmatique, séquentielle et plus généralement encore co(n)textuelle de son objet. Dans ce cadre, nos principaux travaux insistent classiquement sur le traitement des co-occurrences [Mayaffre 2008a, 2008b, 2008c, 2012, 2014] car la co-occurrence articule, dans son essence même, un processus sélectif et un processus combinatoire[15]. Plus précisément, le calcul de la co-occurrence a pu y être conceptualisé comme le premier mouvement d'une statistique lexicale contextualisante, faisant passer nos pratiques, écrivions-nous, d'une lexicographie passive (le relevé d'occurrences) à une lexicologie active sémantiquement. Dans la perspective herméneutique qui est la nôtre, nous avons en effet pu définir la co-occurrence comme *la forme minimale du contexte* (forme minimale calculable) nécessaire à l'interprétation. En une phrase simple mais définitive : constater que le mot *a* et le mot *b* sont co-occurrents, c'est contextualiser minimalement l'un par l'autre[16]. Ainsi pouvons-nous prétendre tenir les deux extrémités du *contexte* c'est-à-dire de la ressource interprétative : à un pôle, le corpus réflexif (forme maximale formalisable du contexte), à l'autre pôle la co-occurrence (forme minimale calculable du contexte).

1.5. Corpus et numérique

A vrai dire, assimiler le corpus à un macro-texte à traiter est moins, aujourd'hui, une option théorique ou une vue de l'esprit qu'une possibilité pratique offerte par le numérique. De manière générale, l'ensemble de notre réflexion sur les corpus retracée dans cette contribution se trouve déterminée par le phénomène numérique. Particulièrement, la notion de *corpus réflexif* ne serait qu'une vaine spéculation si le numérique ne rendait pas possible la structuration architextuelle du corpus et une navigation / exploitation hypertextuelle généralisée : ici c'est bien le numérique et lui seul qui organise le dialogue souhaité entre textes, et objective ainsi l'intertextualité.

Après d'autres, nous avons donc écrit que le passage du papier au numérique ne représente pas un simple changement technique de support du vecteur principal de la culture humaine (le texte) mais une révolution culturelle et épistémologique (anthropologique aussi sans doute), sans guère de précédent dans l'histoire ; sans doute supérieur à la révolution Gutenberg de la Renaissance [Mayaffre 2007b]. De MacLuhan à la médiologie de Debray, et aujourd'hui au développement des humanités numériques, tout indique que le média numérique informera fondamentalement la forme, le fond, la signification ou sens des textes à venir.

Pour le simple propos qui nous intéresse ici – le statut des corpus textuels–, cette révolution peut être résumée par une formule paradoxale qui retourne l'état de l'art traditionnel : le numérique *dématérialise le texte* et *matérialise le corpus*.

Là où l'on avait tendance à naturaliser le texte en l'assimilant, dans sa fixité matérielle, à son support physique traditionnel (la page et le livre), la philologie numérique rend évidente son artefactualité. Pluralité des formats et des codages, choix multiples et individuels dans l'affichage, multiplication des niveaux d'étiquetage, d'annotations, d'enrichissement, circulation, sans limite et jusqu'à l'ubiquité, par fichier joint, parcours de lecture variés, etc. : tout se combine pour souligner aujourd'hui la volatilité ou la relativité du texte, sa dimension artefactuelle, conventionnelle ou culturelle (*i.e* non naturelle). En l'arrachant du scriptorium et

[15] Concrètement, il est possible de montrer que le calcul des co-occurrences mobilise / produit une liste paradigmatique (le mot pôle et les mots co-présents) et une fenêtre syntagmatique ou co(n)textuelle dans laquelle ces mots se combinent.

[16] Après d'innombrables travaux, la pertinence de l'approche n'a plus à être démontrée. En un seul exemple grossier : calculer que la cooccurrence de « classe » est « ouvrière » dans le texte A et « tableau » dans le texte B permet de désambiguïser sémantiquement (et idéologiquement) le mot « classe » et les textes A et B.

de la bibliothèque, en le « défixant » de l'ouvrage papier qui jusqu'ici le supportait, le numérique a définitivement dénaturalisé et dématérialisé le texte, retrouvant ainsi une pratique ancienne de l'Antiquité jusqu'au Moyen-Âge [cf. la réflexion engagée sur la *philologie numérique* par divers auteurs déjà cités : Rastier 2001, chap. III *Philologie numérique* 73-97 ; Viprey 2005][17]. Dans les mots de Dominique Legallois :

> Bien sûr, il faut mentionner l'hypertextualité liée à la mutation numérique du texte, qui oblige à une reconsidération de l'unité textuelle et du texte lui-même : en tant qu'ensemble de possibilités de parcours, le texte devient alors *une unité virtuelle* » (souligné par nous). [Legallois 2006 : 7]

Inversement, là où l'on considérait le corpus seulement comme une idée ou une virtualité, le numérique le matérialise, l'incarne, le réifie en le rendant, quelle que soit sa longueur, palpable et manipulable, exploitable et réexploitable, archivable et échangeable[18]. Si le terme de corpus avait tendu à disparaître dans les années 1980, avant de s'imposer aujourd'hui, c'est qu'il était sans grande pertinence et sans contrainte ; flasque jusqu'ici, il est devenu désormais un concept dur. Le corpus était en effet un idéal (« potentiellement tous les textes susceptibles de m'intéresser ») : c'est aujourd'hui un matériau (« réellement, seuls les textes que j'ai ***saisis*** en machine et que je peux matériellement soumettre au traitement »). Hier encore horizon, il est devenu aujourd'hui, à la faveur du numérique, un continent, dont la clôture constitue une limite mais sur lequel il est désormais possible de circuler. En ce sens, rappelons que le développement de la linguistique de corpus (c'est-à-dire le dépassement du texte seul par le corpus, considéré alors comme le macro-objet de la linguistique) est un produit de la révolution numérique. Du *Brown corpus* au *British national corpus* en passant par le *Trésor de la langue française*, des corpus lemmatisés du Lasla aux corpus XMLisés de la *Base de Français Médiéval* en passant par le *Nouveau corpus d'Amsterdam* – sans rien dire des corpus particuliers des chercheurs –, tous les linguistes de corpus travaillent aujourd'hui sur corpus numériques, et tous réfléchissent à des méthodes numériques pour traiter leur objet numérique.

Récemment, le traitement des corpus numériques par l'Intelligence artificielle et le *deep learning*, nous a permis de montrer comment des notions aussi fuyantes que l'interdiscours ou l'intertexte pouvaient être, pour la première fois, formalisées grâce au numérique. Après apprentissage, la machine classe et décrit le parler de Macron, de de Gaulle, de Pompidou ou de Hollande, et retrouve automatiquement les observables linguistiques qui traversent le corpus présidentiel sous la Vème République : l'intertexte commun que les présidents partagent (Mayaffre et al. 2020a, et 2020b).

Conclusion

Dans un article récent, dans lequel l'auteur fait référence au sein d'une riche bibliographie internationale aux travaux de François Rastier ou de la revue *Corpus* et, plus modestement, aux nôtres, [Laks 2008] rappelle l'opposition fondamentale entre une linguistique du *datum* et une linguistique de l'*exemplum*. Magistralement, Bernard Laks démontre que les corpus existent depuis toujours en philologie et en linguistique, et qu'il ne saurait y avoir de

[17] Cf. aussi la réflexion de [Legallois 2006 : 7] : « Bien sûr, il faut mentionner l'hypertextualité liée à la mutation numérique du texte, qui oblige à une reconsidération de l'unité textuelle et du texte lui-même : en tant qu'ensemble de possibilités de parcours, le texte devient alors *une unité virtuelle* » (souligné par nous).

[18] Nous avons souligné supra la difficulté d'archiver sur le long terme les corpus d'étude, mais le temps de la recherche le corpus numérique peut être stocké, mobilisé et remobilisé, échangé, confronté à d'autres, etc.

modélisation du langage sans prise en compte des usages c'est-à-dire de données attestées recueillies dans de vastes compendiums.

Pour pertinente qu'elle soit dans l'histoire des sciences du langage, l'opposition retracée perd de son efficacité aujourd'hui, en opposant tout le monde à personne et en plaidant une cause entendue désormais par tous ; depuis la fin du XX^{ème} siècle, il n'y a plus nécessité de dénoncer la grammaire générative et ses *exempla* controuvés, comme le temps des luttes politiques contre le rideau de fer apparaît révolu.

Aujourd'hui, ce n'est plus le corpus qui sépare en SHS puisqu'il est admis unanimement. C'est son statut heuristique ou épistémologique qui fait toujours clivage.

Pour Laks, un corpus rassemblerait des données qu'il conviendrait ensuite d'extraire, de décrire, de modéliser. Mais la notion de « données », elle-même, nous semble dangereuse. Les données sont seulement *ce que l'on se donne* ironisent [Malrieu et Rastier 2001 : 554 note] ; elles ne peuvent que difficilement prétendre être le contenu objectif de la langue ou un échantillon représentatif de l'infini du langage. Dangereuse par son illusion objectivante, la notion de « données » semble surtout inopérante pour désigner un *texte* dans son épaisseur et sa complexité, dans son expression que la philologie toujours discute, dans son organisation et sa productivité sémantiques qui restent à découvrir.

Un corpus textuel – puisque tels sont nos corpus – ne donne jamais accès à aucun contenu objectif : il problématise seulement la lecture et organise l'interprétation. Il ne restitue point un sens positif : il le produit. Pas plus que nous pouvons envisager de travailler sans corpus, nous ne pouvons concevoir le corpus seulement comme une chambre froide de dissection du sens ou comme un cercueil de données.

Le corpus est un ensemble dynamique et contrastif pour une sémantique différentielle. C'est un tout vivant, clos sur lui-même mais réflexif pour une herméneutique endogène. C'est une composition matérielle car saisie, contraignante par sa clôture, exigeante par sa réflexivité, mais que le numérique aujourd'hui fertilise et anime.

Il n'y a de sens que d'interprétation : les corpus textuels numériques bien formés sont pour nous le lieu effectif de cette interprétation, et la condition nécessaire de son contrôle.

Références

Adam, Jean-Michel. 2006. « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles ». *Conférence d'ouverture aux JADT 2006* [texte en ligne *Lexicométrica* (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)]

Adam, Jean-Michel et Heidmann Ute (éds.). 2005. *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*. Genève : Slatkine Erudition.

Adam, Jean-Michel. 2008 – éd. revue et augmentée). *La linguistique textuelle. Introduction à l'analyse textuelle des discours*. Paris : Colin.

Aijmer, Karin & Altenberg, Bengt (éds.). 2002. *Advances in Corpus in Corpus Linguistics*. Amsterdam : Rodopi.

Biber, Douglas. 1988. *Variation accross speech and writing*. Cambridge : Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation : A Cross-linguistic Comparison*. Cambridge : Cambridge University Press.

Biber, Douglas. 2004. « Conversation text types : A multi-dimensional analysis », *in* G. Purnelle, C. Fairon, A. Dister (éds), *JADT04*. Louvain : Presses universitaires de Louvain, pp. 15-31.

Biber, Douglas, Conrad Susan & Reppen Randy. 1998. *Corpus linguistics. Investigating language, Structure and Use*. Cambridge : Cambridge University Press.

Brunet, Etienne. 2011. *Ce qui compte. Méthodes statistiques*. Paris : Honoré champion.

Calas, Frédéric. 2006. *Cohérence et discours*. Paris : Pups.

Charolles, Michel. 1983. « Coherence as a principle in the interpretation of discourse ». *Text*, 3-1, pp. 71-99.

Charolles, Michel. 1995. « Cohésion, cohérence et pertinence du discours ». *Travaux de linguistique*, 29, pp. 125-151.

CORPUS, revue en ligne : https://journals.openedition.org/corpus/

CORPORA, peer-reviewed journal of corpus linguistics, Edinburgh Unversity Press.

Détrie, Catherine, Siblot, Paul., Verine, Bertrand. 2001. *Termes et concepts pour l'analyse du discours. Une approche praxématique*. Paris : Champion.

Guaresi, Magali et Mayaffre, Damon, « Intelligence artificielle et discours politique. Quelles plus-values interprétatives ? Application aux corpus parlementaire et présidentiel contemporains » in Damon Mayaffre et Laurent Vanni (éds), *L'intelligence artificielle des textes*, Paris, Champion, 2021, pp. 131-182. [hal-03347997]

Guilhaumou, Jacques. 2006. *Discours et événement. L'histoire langagière des concepts*. Besançon : Presses Universitaires de Franche-Comté.

Guiraud, Pierre. 1954. *Les Caractères statistiques du vocabulaire*. Paris : PUF.

Habert, Benoit, Nazarenko, Adeline. & Salem, André. 1997. *Les linguistiques de corpus*. Paris : Colin.

Halliday, Michael & Hasan, Ruqaiya. 1976. *Cohesion in English*. Londres : Longman.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London & New York : Longman.

Laks, Bernard. 2008. « Pour une phonologie de corpus ». *Journal of French Language Studies*, 18, pp. 3-32.

Lebart, Ludovic, Pincemin, Bénédicte & Poudat, Céline. 2019. *Analyse des Données Textuelles*. Montréal : Presses de L'Université du Québec.

Legallois, Dominique. 2006. « Le texte et le problème de son et ses unités : proposition pour une déclinaison ». *Langages*, 163, pp. 3-9.

Longrée, Dominique & Mellet, Sylvie. 2013. « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours ». *Langages* n° 189, pages 65 à 79.

Maingueneau, Dominique. 1996. *Les termes clés de l'analyse du discours*. Paris : Seuil.

Malrieu, Denise & Rastier, François. 2001. « Genres et variations morphosyntaxiques ». *Traitement Automatique des langues*, vol. 42, n°2, pp. 548-577.

Mayaffre, Damon. 2002. « Les corpus *réflexifs* : entre architextualité et hypertextualité », *Corpus*, n°1, 2002, pp. 51-69. [hal-00554248].

Mayaffre, Damon. 2007a. « L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie/topologie textuelle ». *Lexicométrica*. [hal-00551468]

Mayaffre, Damon. 2007b. « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », François Rastier & Michel Ballabriga (éds), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse : PUT, pp. 15-26. [hal-00551477]

Mayaffre, Damon. 2008a. « Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence », in Serge Heiden et Bénédicte Pincemin (éds.), *JADT 2008, 9es journées internationales d'analyse statistique des données textuelles*, Lyon, Pul, vol. 2, pp. 811-822. [hal-00551300]

Mayaffre, Damon. 2008b. « L'entrelacement lexical des textes, cooccurrences et lexicométrie ». Revue électronique *Texte et corpus*, n°3, 2008, Actes des Journées de la linguistique de Corpus 2007, pp. 91-102. [hal-00553808]

Mayaffre, Damon. 2008c. « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie ». *Sémantique & Syntaxe*, n°9, pp. 53-72. [hal-00551114]

Mayaffre, Damon. 2010. *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*. Thèse HDR, 3 vol. 107, 232, 414 p., soutenue à Nice, le 30 avril 2010. [tel-00655380]

Mayaffre, Damon. 2014. « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », in E. Néé, M. Valette, J.-M. Daube et S. Fleury (ed.), *JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis*,, Paris, Inalco-Sorbonne nouvelle, pp. 15-32. [hal-01181337]

Mayaffre, Damon. 2021. *Le mystère Macron. Ses discours décryptés par la machine*. La Tour d'Aigues : Editions de L'Aube.

Mayaffre, Damon, Pincemin, Bénédicte & Poudat, Céline. 2019. « Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours ». *Langue française*, n°203, pp. 101-115. [hal-02419199]

Mayaffre, Damon *et al*. 2017. « Les mots des candidats, de 'allons' à 'vertu' » *in* Pascal Perrineau (dir.). *Le vote disruptif. Les élections présidentielle et législatives de 2017*, Paris, Presses SciencesPo, pp.129-152 [hal-01635941]

Mayaffre, Damon et al. 2020a. « Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'intelligence artificielle ». 7ème Congrès Mondiale de Linguistique Française.

Mayaffre, Damon *et al*. (sous la dir.). 2020b sous presse. *L'intelligence artificielle des textes*. Paris : Champion.

Mayaffre, Damon et Vanni, Laurent (éds.), *L'intelligence artificielle des textes*, Paris, Champion, 2021

Mayaffre, Damon & Viprey, Jean-Marie. 2012. « La cooccurrence. Du fait statistique au fait textuel : présentation ». *Corpus*, n°11, pp. 7-19. [Revue.org : http://corpus.revues.org/2200]

Mellet, Sylvie & Longrée, Dominique. 2009. « *Syntactical Motifs and Textual Structures. Considerations based on the Study of a Latin historical Corpus*, *in* S. Mellet et D. Longrée, *New approaches in text linguistics.* Amsterdam : John Benjamins, pp. 161-173.

Mellet, Sylvie & Longrée, Dominique (dir.). 2009. *New approaches in text linguistics.* Amsterdam : John Benjamins, pp. 161-173.

Metwally, Heba. 2017. *Les thèmes et le temps dans Le Monde diplomatique (1990-2008)*, thèse de doctorat, Nice 2017

Muller, Charles. 1977. *Principes et méthodes de* statistique *lexicale*. Paris : Hachette.

Partington, Alan. 1998. *Patterns and Meanings : Using Corpora for English Language Research and Teaching*. Amsterdan : John Benjamin's.

Partington, Alan, Morley, John & Haarman, Louann. (éds.). 2004. *Corpora and Discourse*. Proceedings of ComConf 2002 Unversità degli Studi di Camerino, Centro Linguistico d'Ateneo Sept 27th-29th 2002. Berlin : Peter Lang.

Poudat, Céline & Landragin, Frédéric. 2017. *Explorer un corpus textuel. Méthodes – pratiques – outils*. Louvain-la-Neuve et Paris : De Boeck Supérieur.

Rastier, François. 2001. *Arts et sciences du texte*. Paris : Puf.

Rastier, François. 2011. *La mesure et le grain. Sémantique de corpus*. Paris : Champion.

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

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam : John Benjamin's Publishing.

Valette, Mathieu. 2008. « *A* quoi servent les lexiques sémantiques généralistes ? Discussion et propositions ». *Cahiers du Cental*, 5, pp. 43-58.

Viprey, Jean-Marie. 1997. *Dynamique du vocabulaire des Fleurs du mal.* Paris : Honoré Champion.

Viprey, Jean-Marie. 2004. « Analyse séquencée de la micro-distribution lexicale » *in* G. Purnelle, C. Fairon, A. Dister (éds), *JADT04*. Louvain : Presses universitaires de Louvain, pp. 1165-1175.

Viprey, Jean-Marie. 2005-a. « Philologie numérique et herméneutique intégrative » *in* Adam J.-M. et Heidmann U. (éds.), *Sciences du texte et analyse de discours*. Genève : Slatkine, pp. 51-68.

Viprey, Jean-Marie. 2005-b. « Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus » *in* A. Condamines (dir.), *Sémantique et corpus*. Paris : Lavoisier, pp. 245-276.

Viprey, Jean-Marie. 2006. « Structure non-séquentielle des textes ». *Langages*, 163, pp. 71-85.

Williams, Geoffrey (éd.). 2005. *La linguistique de corpus*. Rennes : PUR.