Le corpus : entre concept, objet et application. Comment percevoir ce champ notionnel pour le jeune chercheur ?

Kheira MERINE
Université d'Oran 2

Abstracts:

The notion of corpus with much debate knows meanings that vary according to the place that is given in the corpus analysis tends to scientificity. Used in many fields, it creates a definitional field that refers to a variety of viewpoints assigning various aspects. In the area of language, the body is either an element confirming the theory by marrying its rules (it is a body support (Mayaffre)), or "a dynamic observed" (ibid) from which models can be later described. The latter type appears to meet the characteristics of macrocorpus on which base the description in corpus linguistics. It is immersed in a world where the debates are opposed, but warn that the young researcher may define the notion to make the best use.

Introduction

La notion de corpus chez le jeune chercheur pose ou repose le problème qu'a toujours engendré la relation autonomique entre *théorie* et *pratique.* Si sur le plan épistémologique, la pratique peut modifier la théorie en apportant des éléments nouveaux refaçonnant cette théorie en montrant ses limites, pour le chercheur, il ne s'agira (surtout dans les débuts de sa recherche) que de mener une réflexion pour arriver à un résultat pressenti sur la base d'une analyse menée sur corpus, unique démarche permettant de garantir de la fiabilité des résultats. Alors se pose le problème du choix du corpus, de sa nature, de sa composition par rapport à la (ou aux) variable(s) à traiter et

de sa spécificité qui détermine son rôle dans l'analyse (rôle référentiel, analytique ou de test). C'est dire que la notion de corpus est vaste et peut renvoyer à des aspects divers, d'où la nécessité de cerner cette notion en fonction des acceptions qui lui sont attribuées par les points de vue différents des chercheurs.

1. Comment définir un corpus ?

La notion de *corpus* qui n'est pas tout à fait nouvelle[1], suscite un grand intérêt de la part des chercheurs d'horizons confondus qui essaient de lui attribuer des significations selon le centre d'intérêt de chacun.

Il est vrai que, même si de nombreux domaines se trouvent liés à cette notion, notamment le domaine juridique, socio-anthropologique, philologique..., on a tendance à l'assimiler au domaine linguistique[2], considérant la langue comme matériau de base pour la constitution de tout corpus. Les linguistes ne sont pas toujours d'accord sur la composition du corpus, pour les uns, le corpus est « un vaste ensemble de mots» (Sinclair)[3]ou « un recueil de pièces ou de documents qui concernent une même matière, discipline ou doctrine. » (J-P Dalbéra, 2002)[4], quand ce n'est pas un architexte ou archive de textes[5]

2. Nature du corpus :

---

[1] Habeas corpus

[2] Linguistique de corpus

[3] Cité par F. Rastier : Rastier, François. Enjeux épistémologiques de la linguistique de corpus. *Texto !* [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur :
<http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.

[4] Jean-Philippe Dalbera (2002), « Le corpus entre données, analyse et théorie » in Corpus et recherche linguistique, corpus [en ligne], 1 novembre 2002, mis en ligne le 15décembre2003, Consult2 le 26 septembre 2012
URL : http://corpus revues.org.index 10h .tml

[5] Tel que le conçoit F. Rastier (2004 - 2005)

D'après D. Mayaffre, le corpus peut être soit *un outil qui permet de rendre compte d'une réalité transcendante (la langue ?), d'accéder à un monde déjà-là, d'illustrer une connaissance a priori, de ''découvrir'' un savoir déjà su »*, soit « *un objet vivant de recherche et de connaissance, en lui-même, dont la description débouchera sur des modèles sémantiques à inventer »(Mayaffre, 2005)*
Nous avons-là une définition générale illustrant les cas de figure où on sent la nécessité de recourir à un corpus. Elle nous expose, cependant deux types de corpus :

(i) celui qui nous mène vers le déjà connu, le travail serait donc de confirmer ce qui a déjà été attesté dans d'autres circonstances, ce serait donc un travail mené sur la base d'une théorie, soit pour l'illustrer ou pour garantir la fiabilité de ce corpus qui répondrait à cette théorie.
Dans cette démarche la formation du corpus doit répondre à l'exigence d'une théorie, ce que Mayaffre appelle « l'observatoire d'une théorie » ou alors le »corpus support »

(ii) Celui dont la description va aboutir vers quelque chose de tout à fait nouveau et qui va conditionner le théorique. Ce type de corpus questionne « l'épistémè » de la discipline, il est le « corpus apport ».

3. **Le corpus en linguistique**.

Au départ la notion de « corpus » a divisé le monde de la linguistique en deux parties. La première regroupe les « théoriciens » qui refusaient d'emblée l'existence de corpus dans une démarche linguistique ; parmi eux, il y a surtout les génératives, l'exemple en est cette

déclaration de Chomsky « Corpus linguistics does not exist » (entretien avec Baas Aarts, 1999)[1] La deuxième partie est celle des « descriptivistes » qui ne considéraient aucune théorie sans passer par la description d'un corpus.

Pour les premiers, attachés à la théorie Saussurienne telle que l'a présentée C. Bally en précisant l'objet de la linguistique qui est « la langue envisagée en elle-même et pour elle-même », il n'y a pas lieu d'utiliser un corpus, car tout ce qui provient de corpus, surtout textuels, où il s'agit de performances dans des situations concrètes, relèverait de la sociolinguistique, de la psycholinguistique, de l'analyse du discours ou même de la littérature (Mayaffre, 2005). Ce rejet pourrait s'expliquer par le fait que le fondamentalisme grammatical n'acceptait pas d'être remis en cause par les emplois de l'usage qu'il considérait comme étant des atteintes au système.

Les descriptivistes, appartenant au deuxième groupe, rejettent toute théorie si ce n'est celle qu'ils déterminent à partir d'un corpus ; pour eux, le sens est fixé par l'usage, donc la signification d'une unité linguistique, lexie ou syntagme ne peut être définie qu'à partir d'usages recueillis sous forme de corpus. Leur courant a influencé plusieurs entreprises dont celle du TLF (Trésor de la langue française) qui n'explique le sens d'une entrée lexicale qu'à travers ses différents emplois recueillis dans des romans ou autres écrits littéraires.

Les linguistes appartenant à la position médiane considèrent que la théorie est nécessaire mais pas

[1] F. Rastier (2004)

suffisante pour décrire des comportements langagiers qui souvent sont entachés de singularité.

4. Forme(s) du corpus.

Là aussi les avis se divisent en deux grands groupes : ceux qui reconnaissent une pluralité formelle au corpus, et ceux qui n'y voient qu'une seule forme liée au texte ou au macrotexte.

4.1 Pluralité formelle du corpus

Pour les linguistes qui partagent ce point de vue (Dalbéra,2002, Mayaffre, 2005...), il est aisé de passer à une classification des corpus en fonction de leur formes et composition. C'est ainsi que l'on peut distinguer le corpus oral du corpus écrit, le corpus composé de fragments de langue tels que les phonèmes, morphèmes, lexèmes, syntagmes et autres, du corpus textuels, allant du texte réduit (un roman, une œuvre) à l'ensemble de textes (archive de textes). Il peut être clos ou ouvert, construit ou authentique. C'est dans un souci méthodologique que se fait le choix du corpus.

Concernant cette classification, Mayaffre (ibid) établit une sorte de hiérarchie en présentant trois grandes formes de corpus :

-Le corpus lexicographique dont la spécificité est de *prétendre à l'exhaustivité* Et de ce fait, ils présentent un caractère clos..

-Le corpus phrastique utilisé pour des recherches en syntaxe et morphosyntaxe. Ce corpus peut être recueilli comme il peut être construit (fabriqué).

-Le corpus textuel qui pose le problème du sens. Pour Mayaffre (ibid), ce type de corpus exclut toute idée d'exhaustivité et, offrant plusieurs possibilités

d'interprétations[1], il ne peut permettre une véritable détermination du sens. On peut fabriquer des corpus réduits (mot/phrase) mais pas un texte.

4.2 Vers des corpus à caractère uniforme ?

> Par *uniforme*, nous ne visons pas une standardisation de la forme du corpus, mais une uniformisation des règles à suivre pour le choix et la composition de corpus. Les adeptes de ce point de vue avec, à leur tête, Rastier (2002, 2004, 2005), placent la notion de *corpus* en linguistique au sein d'une discipline appelée « la linguistique de corpus ». Selon eux, la définition du corpus est liée à deux éléments fondamentaux qui sont la *représentativité* et *l'homogénéité* (Rastier,2002)

La représentativité vise l'importance quantitative du corpus (archives de textes, ou architexte (Rastier)), et l'homogénéité, quant à elle, indique la nécessité d'aligner les textes selon les mêmes genres et sous-genres, car « les relations sémantiques s'établissent préférentiellement entre textes du même genre, du même champ générique et du même discours » (ibid, 2004)

Les modalités d'exploitation de ces corpus reposent essentiellement sur l'outil informatique qui, à l'aide de pratiques logométriques (lexicométrie et textométrie) permet un dépouillement efficace du macrocorpus. Ce dépouillement se fait en fonction des *corrélations* qui

[1] C'est pour mieux cerner cette pluralité interprétative qu'est née la sémantique interprétative qui se dote de l'outil informatique dans ses investigations. (pour une meilleure compréhension du phénomène, voir Pincemin, 2012)

existent entre les différents paliers du corpus que Rastier (ibid) présente selon la hiérarchie suivante :

- le microtextuel (morphème, lexie),
- Le mésotextuel (de la période au chapitre),
- Le macrotextuel (texte complet dont péritexte et paratexte),
- L'intertextuel (le corpus).

Ainsi, tout élément servant à la composition du corpus peut être mobilisé pour l'étude en fonction d'un objectif donné. Il faut noter que pour la linguistique de corpus, l'objectif premier est la quête du sens. De nombreux travaux ont été réalisés dans ce cadre et ont démontré que les corrélations entre deux paliers différents peuvent aboutir à une caractérisation ou à une détermination d'éléments qui ne figuraient pas dans le cadre normatif

Comme exemples, on propose deux résultats, l'un pris des travaux de Bourion (2001, pp.42-45), et l'autre des travaux de Nathalie Deza (1999) tous deux cités par Rastier (ibid)

Pour le premier (celui de Bourion), le travail a porté sur les expressions « au pied de » et « aux pieds de ». Le corpus pris de la banque Frantext (un nombre très important de textes) a révélé, après analyse, deux réalités. La première est que « au pied de » (singulier) renvoie à l'expression d'une verticalité (au pied d'un arbre, au pied de la montagne…), alors que « aux pieds de » (pluriel) renvoie à une scène où l'on prie, implore quelqu'un. Conclusion,

l'expression au singulier donne lieu à une *localisation* alors qu'au pluriel, elle devient une *configuration narrative*. On en a déduit que chacune des deux expressions nécessite une entrée (à part) dans le dictionnaire. Ce travail fait état d'une corrélation entre « lexique » et « genre ».

Le deuxième est l'aboutissement d'un travail mené sur le roman français de 1830 à 1970. Le corpus est composé de 350 œuvres dans lesquelles sont répertoriées 4488 mentions d'âge des personnages. Après des calculs et des comparaisons, la chercheure aboutit au résultat suivant ; les âges qui reviennent le plus souvent (*sur représentés*) sont 16, 18 et 20 ans. La conclusion à en tirer c'est que dans le roman français, on a presque toujours 20 ans. Dans ce travail, on s'est basé sur deux variables textuelles : le texte et l'intertexte pour étudier la *stéréotypie* et *la norme de la doxa* représentées dans les écritures romanesques.

D'autres résultats sont obtenus, concernant la corrélation entre d'autres paliers[1] et confirmant la thèse des défendeurs de la linguistique de corpus, pour qui le global (macrocorpus) détermine le local (stucture minimale) et non l'inverse, autrement dit dans un langage plus simple : la phrase détermine (sémantiquement) le mot, le texte détermine la phrase, et le corpus, le texte. A cet effet, Rastier (ibid) précise : « la corrélation entre descriptions locales et description globale permet de préciser l'articulation entre la problématique du signe et la

[1] Voir Rastier, 2002-2004.

problématique du texte, en subordonnant la première à la seconde ».

Avec de tels travaux et de tels résultats on aboutit à deux comportements nouveaux : (i) on enrichit ou développe les typologies déjà existantes mises en place par la norme. (ii) on ajoute d'autres typologies non prises en ligne de compte par la norme, en travaillant aussi bien sur *l'invariance* que sur *la singularité* de la langue..

4.1.1. Linguistique de corpus et norme

Ce rapport s'explique par l'évolution qu'a connue la linguistique de corpus. En effet, celle-ci s'est trouvée réconfortée par le tournant théorique qu'a connu la linguistique générale après les révélations concernant la manière dont Bally a limité l'objet de la linguistique (voir §3). C'est ainsi que, dans un manuscrit découvert récemment, Saussure estime que

« l'entreprise de classer les faits d'une langue se trouve donc devant ce problème : de classer des accouplements d'objets hétérogènes (signes-idées), nullement, comme on est porté à le supposer, de classer des objets simples et homogènes, ce qui serait le cas si on avait à classer des signes ou des idées. Il y a deux grammaires, dont l'une est partie de l'idée, et l'autre du signe ; elles sont fausses ou incomplètes toutes deux. » (2002 : 20).(cité par Rastier, ibid)

Ainsi est remis en cause l'aspect *homogène* attribué à la langue considérée comme système devant se baser, exclusivement, sur l'étude du signe, placé au centre de toute analyse. Cette remise en question d'une théorie qui a longtemps régné sur toutes les approches linguistiques va entrainer une autre concernant un deuxième fait saussurien, à savoir la dichotomie : langue/parole.

La dichotomie langue/parole (discours) est considérée plutôt comme une « dyade » (Mayaffre, 2005) où dans l'opposition, il y a complémentation. Pour Rastier (2002-2004), cette distinction (entre langue et discours (parole)) donne lieu à une antinomie dont l'articulation, n'a été expliquée par aucune théorie à commencer par l'approche aristotélicienne qui, opposant l'acte à la puissance, donnait la primauté à la puissance sur l'acte, elle est suivie de la position de Humboldt qui oppose : energeia et ergon reprise par Chomsky sous le couple de compétence/performance. Or précise- t-il *la langue ne préexiste pas à la parole : elle est apprise en son sein, et la compétence des sujets évolue au cours de leurs pratiques effectives. (ibid).* Coseriu (1969), quant à lui, explique que « *le chainon manquant entre langue et parole est constitué par l'espace des normes* » (Rastier, ibid) A travers ces débats, nous comprenons d'abord que le mode d'analyse ne doit pas être freiné par les limites de la norme, bien au contraire, il s'agit « d'inverser la donne » et d' « exploiter les corpus pour décrire la norme » (Rastier, ibid)

Ces principes vont être concrétisés à travers des travaux dont on a présenté quelques exemples, mais les adeptes de cette linguistique projettent la développer à travers des pratiques descriptives telles que celles de la sémantique interprétative où il sera question d'interroger la sémiosis du texte.

A ce niveau, nous remarquons que ce type de corpus (architexte, macrocorpus) devient un objet mis au service de l'épistémologie, il représente bel et bien un corpus apport.

5. Corpus : du concept à l'objet

Quelle que soit la conception que l'on a du corpus – *sac de mots* ou archive de textes – (Rastier, 2002), celui-ci est, d'après Mayaffre (2005) un objet heuristique, arbitraire n'ayant de sens que par rapport à l'objet de l'étude pour laquelle il a été construit. Un même corpus peut être traité dans des domaines différents, car cela dépendra des questions qui lui sont posées, des réponses attendues et des résultats escomptés. Exemple : un architexte peut intéresser un historien, un philosophe, un philologue, un linguiste, un littéraire, etc. Ainsi, n'utilise-t-on un corpus que pour arriver à un résultat (pressenti ou non), pour aller vers l'inconnu, qui doit devenir connu.

A ce propos Mayaffre (2005) précise : « Ce n'est pas un donné disciplinaire mais un objet heuristique. Le contenu objectif ou matériel d'un corpus textuel n'appartient pas à l'Histoire, à la Linguistique ou à la Philosophie. C'est l'intention du chercheur qui est importante et lui donne son sens ».

Exemple : l'emploi récurrent d'un lexique de guerre chez Alphonse Allais surtout dans « Royal Cambouis » peut être interprété comme révélateur d'une époque bien dominée par la guerre de 1914-1918 (maréchal des logis, tringlot, train des équipages...) et donner lieu à une étude psychologique en vue de voir comment on évacue le stress de la guerre par le rire. Mais cela pourrait être aussi un bon corpus pour un linguiste qui travaillerait sur les archaïsmes

C'est pourquoi dans un corpus tout est méthodologie avec précision de la démarche et de l'objectif comme le souligne Mayaffre (ibid), dans le passage suivant : « L'heuristique s'interroge avant tout, dès l'origine, sur les techniques pour extraire des résultats. C'est, semble-t-il, la préoccupation première d'Aristote dans l'*Organon*. La découverte passe par des *techniques* heuristiques, par un *ars inveniendi*, à terme, dans l'évolution des sciences, par une *méthode* constituée de traitement. Il faut un protocole méthodologique – une procédure intellectuelle et des procédés techniques explicites – pour traiter un corpus. » Le lien entre le corpus et la méthode répond à une dialectique que Mayaffre (ibid) explique dans ces mots « le corpus commande la méthode et la méthode ordonne le corpus. »

Conclusion

Dans notre réflexion, nous avons essayé de cerner d'une manière très théorique la notion de corpus en nous basant sur divers travaux, notamment sur des réflexions exposées sous forme d'articles appartenant surtout à F. Rastier, D. Mayaffre et P. Dalbéra. Ces réflexions nous ont permis de circonscrire l'objet de notre réflexion par rapport au corpus dans le domaine de la linguistique, puis d'en apporter des

classifications avec chacune ses caractéristiques. Ces classifications sont conçues pour avertir de la technique d'exploitation à adopter au corpus, en fonction de son type et de la variable à étudier.

Bibliographie :

Condamines A. et al. (1999). Corpus et traitement automatique des langues : pour une réflexion méthodologique, Actes de l'atelier thématique TALN, Cargèse.

Dalbera, Jean-Philippe « Le corpus entre données, analyse et théorie », *Corpus* [En ligne], 1 | novembre 2002, mis en ligne le 15 décembre 2003, Consulté le 26 septembre 2012. URL : http://corpus.revues.org/index10.html

Guilhaumou, Jacques , « Damon Mayaffre — Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République. Paris : Champion, 2004, 292 pages (50 €). », *Corpus* [En ligne], 4 | décembre 2005, mis en ligne le 05 septembre 2006, Consulté le 27 mai 2012. URL : http://corpus.revues.org/index322.html

Mayaffre Damon. (2000). Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres. Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939). Paris : Honoré Champion.

MERINE Kheira,
Maitre de Conférences,
Université d'Oran 2,
Domaine de recherche : Sciences du Langage
Merinekheira2@yahoo.fr