**Master « Linguistique »**

**K. Mérine**

 **Le corpus : objet, concept et théorie**

**Section 3**

**1. Compétence :** Rentabiliser un corpus à partir de variables explicatives ou justificatives de certains phénomènes linguistiques et/ou autres

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 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 (op.cit.)

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 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 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 (op.cit.) précise : « la corrélation entre descriptions locales et description globale permet de préciser l’articulation entre la problématique du signe et la 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..

4.2.1. 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.