EN GUISE D'INTRODUCTION

Le présent volume, le XIX[e], de la revue *Acta Universitatis Carolinae – Philologica – Romanistica Pragensia* est monothématique : c'est l'utilisation de corpus linguistiques qui fait, sur le plan méthodologique, son unité.

Conçus dans un premier temps comme un outil méthodologique somme toute auxiliaire, les corpus de langue sont désormais devenus un élément fondamental et, dans maints domaines, indispensable pour la recherche en linguistique. D'immenses stocks de données permettant, au fur et à mesure, une recherche de plus en plus affinée et confortable, ont sensiblement réduit l'espace laissé à l'intuition du chercheur qui avait constitué, avant l'apparition des corpus, le critère décisif dans l'analyse des données de langue. Les corpus représentent une contribution de taille pour les chercheurs travaillant sur les langues dont ils ne sont pas les locuteurs natifs, mais même un linguiste qui travaille sur sa langue maternelle doit chaque jour se rendre à l'évidence que sa vision apparemment inébranlable de celle-ci ne résiste guère à l'examen fait à partir d'un vaste stock de données provenant d'une grande quantité de locuteurs natifs.

Forte d'une longue tradition, la linguistique de corpus tchèque compte parmi les plus importantes d'Europe. Ceci est dû surtout aux activités de l'Institut de corpus national tchèque auprès de la Faculté des Lettres de l'Université Charles de Prague. Le corpus national tchèque constitué par les soins de cet Institut consiste aujourd'hui d'un certain nombre de sous-corpus contenant une énorme quantité de données. La construction des corpus n'aboutit pas seulement à un accroissement progressif de leur ampleur, mais entraîne logiquement un remaquable développement de la méthodologie elle-même, et ceci sur plusieurs aspects. En décrivant en ce sens la situation de l'école tchèque, on obtient, à une échelle réduite, une image du développement de la linguistique de corpus au niveau international, ce qui fait que tout ce qu'on dira sur les corpus tchèques vaut également pour la linguistique de corpus en général. Un des traits marquants est la mise en place des sous-corpus spécialisés. On assiste à la construction de corpus synchroniques aussi bien que diachroniques, de corpus de langue écrite et de langue parlée, de corpus de registres de langue partiels, de corpus monolingues et plurilingues, etc. La constitution de tels corpus requiert un perfectionnement de l'infrastructure technique : la structuration interne des corpus ne cesse en effet de s'améliorer, ce qui permet à l'usager de cadrer, avec de plus en plus de précision, le choix d'un segment de corpus pour sa recherche. En même temps, on voit s'affiner les outils de recherche qui réduisent la quantité de travail manuel nécessaire

pour l'interprétation des données (cette phase reste et restera à jamais l'élément clé de l'analyse de corpus, mais un usager averti, maîtrisant tous les outils que les corpus mettent à sa disposition, est capable de simplifier sensiblement cette tâche).

L'apparition de corpus parallèles constitue un aboutissement notable des efforts visant à perfectionner la méthodologie et la recherche de type méthodologique. On connaît dans le monde plusieurs corpus parallèles. C'est au sein du Corpus national tchèque qu'on a mis en place le corpus synchronique parallèle *Intercorp* (http://www.korpus.cz/intercorp/) dont la première version a vu le jour en 2008. Au mois de juin 2012, sa cinquième version intègre 27 langues plus le tchèque. Le nombre de mots (à l'exception du tchèque) dépasse 90 millions. Le corpus est conçu de telle manière que le texte de base est en tchèque et les textes identiques dans les différentes langues lui sont reliés parallèlement. Ce qui a un intérêt particulier pour le présent volume, c'est que parmi les 27 langues, on trouve le français, l'italien, le portugais et l'espagnol. Il existe en plus un sous-corpus roumain et un sous-corpus catalan est en ce moment en voie de préparation.

Le lien entre les corpus que l'on vient de mentionner, notamment le corpus parallèle *Intercorp*, et le présent numéro de *Romanistica Pragensia* est de double nature : d'un côté, ce sont les enseignants-chercheurs de l'Institut d'Etudes romanes, c'est-à-dire du département responsable de l'édition de la revue, qui dirigent les différentes sections romanes du projet *Intercorp,* de l'autre côté il y a la méthodologie de la recherche qui constitue en effet une des orientations clés de cet institut. Tout cela se traduit entre autres dans un vaste projet pluriannuel intitulé *Les langues romanes à la lumière des corpus de langue,* réalisé dans le cadre du sous-projet *Linguistique* rattaché au *Programme de développement de la recherche scientifique à l'Université Charles.* Le présent numéro est un résultat de ce projet.

Comme on peut le constater, les différentes contributions sont reliées précisément par la méthodologie de corpus. Elles sont toutes à ranger dans la catégorie qu'on appelle *corpus-based* et non pas *corpus-driven* : la méthodologie de corpus et l'analyse des données de corpus servent à vérifier une hypothèse linguistique préalable. Elles n'ont pas donc l'ambition de développer la méthodologie générale de corpus comme c'est le cas de nombreuses publications éditées par l'Institut de corpus national tchèque, mais bien plutôt d'apporter, sur la base des données de corpus, une contribution à une meilleure connaissances des langues romanes.

S'appuyant sur un même critère méthodologique, les différentes constributions présentent une assez grande variation typologique. On y trouvera des textes basés sur des corpus monolingues, le plus souvent de grands corpus nationaux d'une langue donnée, mais toute une série d'articles se proposent de vérifier la toute récente méthodologie de corpus parallèles. En analysant les données obtenues grâce à *Intercorp,* ils offrent des études contrastives entre le tchèque et les langue romanes ou encore entre les langues romanes. Au-delà de proposer une optique nouvelle pour l'étude de certains phénomènes, ces contributions testent également les possibilités des corpus parallèles pour l'analyse linguistique et vérifient leur utilité en vue de la description des différents niveau de langue.

Si l'on considère les articles réunis ici du point de vue romanistique, on y trouve des contributions consacrées au français, à l'italien, au portugais et à l'espagnol. Certaines

d'entre elles (on vient de mentionner l'approche constrastive de quelques-uns de ces textes) ont une portée romanistique plus générale. Leur variété concerne également les disciplines linguistiques : le lecteur y trouvera des textes (dont certains ont une dimension diachronique) consacrés à la phonétique, la morphologie, la syntaxe et la formation des mots.

Le présent numéro se propose d'apporter une contribution à une étude systématique des langues romanes (non seulement dans une optique contrastive) à l'aide de la méthodologie de corpus. Dans le cadre du projet que l'on a déjà cité, il ne s'agit bien entendu que d'un premier pas qui sera suivi d'autres, focalisés plus spécifiquement sur un certain nombre de sujets de recherche en linguistique romane.

Petr Čermák
Jaroslav Štichauer