Rapport pour soutenance de la thèse de doctorat en co-tutelle de

Monsieur Martin SVÁŠEK
intitulée

" Fratchèque.
Un corpus parallèle bidirectionnel
français – tchèque et tchèque – français
Définition, élaboration, exploitation"

Le document remis pour soutenance se compose d'un mémoire d'environ 250 pages et d'un CD contenant la partie corpus et les solutions informatiques d'analyse des particules et du discours direct que M. Martin Svášek a lui-même élaboré.

Disons immédiatement que la procédure de co-tutelle est parfaitement réalisée dans la mesure où il s'agit réellement d'un double doctorat. En effet, tant l'exposé sur les particules en tchèque que celui sur les particules en français représente chacun un véritable travail scientifique au niveau de ce qui peut être produit par un chercheur autochtone. Disons aussi que M. Svášek ne peut pas être décelé à la lecture de son mémoire entièrement rédigé en français comme étranger. Il s'agit d'un texte parfaitement écrit, dans une langue soutenue sur un sujet difficile tant au niveau de la linguistique que de l'informatique et pratiquement exempt de fautes, ce qui est loin d'être toujours le cas chez des Français de naissance. Cette thèse met ainsi en valeur la parfaite double culture de M. Svášek.

Il convient de mettre également en relief la triple teneur de cette thèse appartenant de plein droit au domaine de la linguistique générale, au domaine des études tchèques et même françaises ainsi qu'à une informatique traitant des problèmes de la plus grande actualité.

La durée actuelle de la réalisation des thèses et particulièrement de celles réalisées en cotutelle est réduite à trois années. Le plan de travail et surtout de financement d'une cotutelle n'est prévu que pour ces trois années quasiment définitoires.

Dans ces conditions, la simple conception et réalisation par un homme seul d'un corpus aligné d'un million de mots dans chacune des deux langues pourrait être considérée comme totalement suffisante, puisqu'il s'agit très généralement du travail d'équipes entières de laboratoires bien dotés. En ce qui concerne la réalisation de ce corpus, M. Svášek a été confronté au choix difficile de textes contemporains significatifs, libres de droit et traduits dans l'autre langue. De ce point de vue, l'intégration, très bienvenue, du présent travail dans les projets bilingues du ČNK a permis au candidat de pouvoir utiliser quelques textes tchèques déjà saisis. Il n'en reste pas moins que M. Svášek a dû acquérir la culture nécessaire à la définition et à la réalisation de corpus, en particulier, bilingues alignés, puis saisir une masse considérable de documents, vérifier intégralement leur correction après saisie optique (un très gros travail fastidieux) et enfin construire le parallèlisme des deux corpus à l'aide du logiciel ParaConc, ce qui n'est pas une mince entreprise! Cette réalisation a été chronophage et a largement empiété sur le temps de réalisation de la thèse.

Au-delà de la création de ce corpus aligné, contribution non négligeable aux projets nationaux tchèques de corpus, M. Svášek a conduit une recherche linguistique d'un très grand intérêt sur les particules, catégorie floue, mal définie et dont on peut dire qu'elle ne constitue vraisemblablement pas une catégorie lexicale, en tout cas pas dans la définition et la fonction que peuvent jouer les catégories habituelles de verbe, substantif, ... , voire même de préposition et conjonction. Il s'agit plutôt dans le cas des particules et dans la mesure où on peut les définir (et M. Svášek montre très clairement qu'elles ont une existence bien plus évidente en allemand et en tchèque qu'en français, où elles apparaissent sous forme de « particules énonciatives ») d'un élément appartenant à la tectonique d'un texte ou d'un énoncé en général, qui, de plus, explicite – au moins dans le cas des particules choisies „vždyť" et „přece" – des relations situationnelles entre locuteur et récepteur détenant tous les deux (du moins en principe) la référence à un savoir non dit.

Le travail fournit sur les particules „vždyť" et „přece" a pour caractéristique essentielle une démarche de science expérimentale rendue possible par l'existence du corpus réalisé par M. Svášek et quelques autres corpus élaborés par ČNK. Ces corpus sont analysés automatiquement à l'aide de deux scripts différents, réalisés à l'aide du langage de programmation Python dont le candidat a assimilé les fonctionnements essentiels. L'un de ces deux scripts est dédié au traitement de l'ensemble des corpus et l'autre au traitement du corpus Fratchèque, le corpus de M. Svášek. Ce corpus montre et respecte, à notre plus grande satisfaction, l'absolue nécessité de parfaitement respecter la typographie qui participe à l'analyse des textes, tant humaine qu'automatique. Seul ce respect de la typographie permet une analyse correcte du discours direct marqué par les guillemets.

Après les exposés sur les „částice" tchèques et les « particules énonciatives » françaises, dont nous avons dit en introduction qu'ils constituent chacun le centre d'une thèse, la partie fondamentale de la thèse de M. Svášek est représentée par la partie consacrée à l'étude empirique parallèle des particules. M. Svášek y analyse les particules „přece" et „vždyť" ainsi que leurs variantes („přec" (caractérisée comme littéraire) et „přeci" d'une part, „dyť" et „dyk", d'autre part) et les collocations „přece jen" et „přece jenom". La présentation étymologique de ces particules est très éclairante. Au-delà des chiffres réels des dépouillements manuels et des résultats d'analyse automatique, M. Svášek fait des projections sur les grands corpus à l'aide d'un coefficient dont il montre la détermination. Statistiques et diagrammes fournis accompagnent l'exposé. La distinction, faite par programme, entre éléments du discours direct et discours indirect ainsi que la correlation entre les particules étudiées et le discours direct sont des résultats très intéressants.

L'étude fine et précise des divers emplois et significations des particules permet à M. Svášek de présenter des applications lexicographiques, an particulier de réalisation de dictionnaires, mais aussi à l'aide des études contrastives avec le français,

d'offrir la matière à des applications pédagogiques pour l'apprentissage du tchèque et du français par des Tchécophones, du tchèque par des Francophones.

Ce que M. Svášek ne souligne pas assez, c'est l'importance de ses recherches pour le développement ultérieur d'analyses automatiques syntaxiques et textuelles du tchèque.

En conclusion, nous sommes en présence d'un travail remarquablement bien pensé où M. Svášek fait preuve de connaissances démontrant son aptitude à l'analyse linguistique pouvant aller dans le détail et sa capacité de définir un système informatique avec la clairvoyance de la structure générale nécessaire.

L'ensemble des travaux est d'un excellent niveau qui classe désormais M. Svášek parmi les spécialistes du domaine. Je recommande donc une soutenance dans les meilleurs délais.

Paris, le 9 octobre 2007

Patrice Pognan
Professeur à l'INALCO
co-directeur de thèse