UNIVERSITE PARIS III
SORBONNE NOUVELLE

Pré-rapport de Jean Léonard Léonard (UMR 7018, Maître de conférences HDR, Université de Paris 3, ILPGA) sur la thèse NR en co-tutelle (Université Charles, Prague-Inalco, Paris) de Martin Svášek : *Fratchèque. Un corpus parallèle bi-directionnel français-tchèque tchèque-français : définition, élaboration, exploitation*, 225 pages, dir. Vladimír Petkevič et Patrice Pognan.

Cette thèse de 207 pages de dissertation doctorale se conforme à une tendance aujourd'hui de plus en plus répandue, qui consiste à traiter un sujet de recherche de manière succincte, plutôt qu'une approche cumulative, fondée sur la confrontation des détails d'une argumentation grammaticale ou philologique. Cette démarche d'ingénieur privilégiant la stratégie rédactionnelle de la « thèse courte » est ici pertinente, dans la mesure où l'essor des recherches exploratrices dans le domaine du T.A.L. (Traitement Automatique du Langage) requiert une mise à disposition rapide et dynamique des résultats de l'application d'un protocole de recherche. Dans la mesure où l'on attend d'un candidat au doctorat d'atteindre trois objectifs – construction de l'objet, analyse critique de la littérature existante, questionnement des méthodes et des procédures –, c'est sur ces trois aspects fondamentaux que nous jugerons ce travail, en tentant de tirer le meilleur parti du produit final de la recherche. Or, dans ces trois domaines, la thèse de Martin Svášek atteint pleinement ses objectifs, et représente une contribution précieuse tant en linguistique générale qu'en T.A.L. Qu'il soit dit également quant aux questions de présentation et de correction stylistique que cette thèse est rédigée de manière tout à fait correcte et élégante, à quelques scories près, tout à fait excusables pour un locuteur-rédacteur non natif en français. On ne peut même que féliciter le candidat pour s'être fait apparemment suffisamment relire et corriger par des locuteurs natifs. La quantité d'erreurs de forme (coquilles et bévues stylistiques) est négligeable, mettant d'autant plus en valeur la très grande qualité du contenu, du point de vue méthodologique, empirique et théorique – les trois grandes qualités de cette thèse. La bibliographie citée est suffisante et à jour dans l'ensemble, à quelques lacunes près concernant les études de corpora de français oral (travaux du G.A.R.S. d'Aix-en-Provence, notamment). La thèse est accompagnée d'un CD-ROM contenant les textes tchèques et français, originaux et traduits, qui ont alimenté la base de données Fratchèque constituée par l'auteur.

Le présent rapporteur interviendra en tant que linguiste généraliste et descriptiviste s'intéressant à des questions de typologie grammaticale. Le questionnement de l'objet « particules » par Martin Svášek lui a rappelé un travail analogue qu'il a dû réaliser il y a quelques années au sujet de la notion de « cas » en grammaire moderne. Tout comme le « cas », les « particules » semblent bien être une sorte de prisme épistémologique balayant de son rayon lumineux tous les champs des composantes formelles de la grammaire, depuis les conditions d'épellation phonologique (l'accentuation et l'intonation, dans le cas des particules) jusqu'aux conditions de spécification discursive, pragmatique et logique, en passant par la structuration morphologique et le balisage syntaxique. Dans cet ordre d'idée, la différence entre ces deux « zones grises » ou « Twilight Zones » de la tradition grammaticale occidentale moderne tient dans une antinomie en termes de marquage flexionnel, puisque le

cas se définit avant tout par son expression désinentielle en surface, tandis que les particules, bien au contraire, se définissent avant tout comme éléments non fléchis, invariables, mais tout aussi polyvalents et répartis dans toutes les composantes formelles que les cas. Ce n'est pas le moindre mérite de cette thèse que d'offrir au linguiste ouvert à divers domaines de langues de types très différents, des pistes explicatives pour des phénomènes morphologiques encore obscurs dans des langues bien éloignées du tchèque et du français. On pense notamment à l'intérêt de cette thèse, tant sur le plan de l'épistémologie critique de la notion de « particules » que du point de vue de la méthodologie de traitement des données dans de vastes corpora parallèles, pour examiner d'un regard nouveau les particules du japonais ou du coréen, voire nombre de morphèmes semi-affixaux de modalités médiatives dans les langues d'Amazonie – même si Martin Svášek , bien entendu, évitant résolument de s'aventurer hors de son domaine de spécialité, ne fait à aucun moment allusion à ces langues. Mais le potentiel est indéniablement sensible dans ce travail exploratoire.

La thèse se présente sous la forme d'un ampoule, dont les deux chapitres d'introduction méthodologique A et B constituent la douille, tandis que la partie C, le corps central, se déploie en une poche contenant l'essentiel de la recherche et de la réflexion théorique en grammaire et en T.A.L.

Après une définition des corpora utilisés par l'auteur, en particulier le corpus *Fratchèque* (pp. 6-12), dont il est le concepteur, le créateur et l'utilisateur, et un survol de la typologie des corpora bilingues existant (*corpus parallèle, de traductions libres, de textes comparables*, etc.) richement documentée en références Internet (p. 9) en partie A, 25 pages sont consacrées dans la partie B à l'élaboration des textes destinés à alimenter *Fratchèque*, y compris à des détails pratiques tels que les techniques de scannage, les problèmes d'alignement des corpora parallèles, d'erreurs de lecture optique. Ces considérations, certes prosaïques mais indispensables, sont accompagnées de tableaux clairement organisés (liste des textes littéraires et des traductions français-tchèque-français retenus par l'auteur, p. 16 et 26; liste d'erreurs de lecture optique, p. 38). La partie C constitue l'essentiel de la thèse (pp. 42-198), où se déploie l'essentiel du travail de linguiste et de programmeur du candidat. Cette partie se ramifie en 6 sections, qui sont en fait des chapitres de taille variable (entre 10 et 20 pages), dont les trois premières sont consacrées à la définition de l'objet – les particules ou *částice* –, d'abord de manière générale, puis dans les traditions tchèque et française de description grammaticale, et les trois dernières, à l'étude de cas servant de base empirique et expérimentale au candidat : ***vždyť*** et ***přece***. Le titre générique attribué à cette partie centrale « Exploitation de Fratchèque » cache en fait une étude comparative intéressante et originale sur les acquis, les hypothèses et les apories sur une classe bien particulière de « parties du discours » que sont ces mots-outils ou foncteurs regroupés sous le terme parapluie de « particules ».

Le chapitre 3.1. traite de la diversité terminologique de l'objet : *particules, mots-outils, connecteurs, particules énonciatives* (PEN de Jocelyne Fernandez-Vest), *mots du discours* (Oswald Ducrot, René Metrich); *particles, Focus Particles, Modal Particles, Sentence Equivalents, Discourse Markers ; částice, Partikeln, Funktionswort, Formwort, Füllwort*, etc. en tentant de délimiter la part de lexique, la part de fonction ou de relation syntagmatique et la part de sémantique, de logique ou de discours qui préside à ces différentes dénominations. Le chapitre 3.2 prolonge cette réflexion liminaire en explorant les propositions des grammairiens modernes tchèques, qui ont tour à tour proposé une taxinomie fondée sur a) le contenu ou l'expressivité, b) le caractère non flexionnel et la distribution ou la position syntagmatique, c) la lexicalité versus la fonctionnalité – les particules relèvent alors de la catégorie des *fonctèmes*, notamment de modalité –, d) d'iconicité dénominative ou imitative *versus* relations de substances ou d'énonciation, e) d'autosémantisme versus synsémantisme, f) de traits adverbiaux, conjonctifs, onomatopéiques, particulaires, g)

d'endocentricité *versus* exocentricité, h) de plans ou strates analytiques, comme le plan gnoséologique, des significations et de la polarité thème-rhème avec une fonction de désambiguation, i) de plan élémentaire hors contexte *versus* plan contextuel *versus* accentuel, j) de traits pragmatiques liés aux expressions illocutives, allocutives et perlocutives, k) de hiérarchisation entre plan de description et de structure communicative, l) de structure informationnelle, m) de modalité et de modalisation, n) de processus de rhématisation, o) de connexité et d'embrayage discursif, etc. En peu de pages, Martin Svášek nous présente un inventaire intelligent et clair de ces différents points de vue, qui construisent un véritable prisme épistémologique. Il en ressort clairement à quel point la tradition de pensée tchèque est, à cet égard, charpentée solidement sur les facteurs et les fonctions du célèbre schéma de la communication de Roman Jakobson, même si l'auteur n'y fait pas directement allusion. Mais les notions de centrage sur la dyade fondamentale allocuteur-allocutaire, le message, et les fonctions émotive et connative, scandent littéralement l'analyse séquencée des différentes propositions théoriques et terminologiques des linguistes tchèques tout au long du XXe siècle. En revanche, la tradition française, présentée au chapitre 3, se caractérise par essentiellement deux approches, qu'on pourrait grossièrement résumer en une approche adverbialiste – celle de la grammaire scolaire, ou « grammaire traditionnelle », et une approche énonciativiste, représentée par Oswald Ducrot et par Jocelyne Fernandez-Vest – l'une dans une perspective discursive et sémantico-logique, l'autre dans une perspective de structure informationnelle, davantage morphosyntaxique et soucieuse de la construction de l'oralité. Martin Svášek n'oublie pas non plus l'école du « conversationnalisme genevois », mais on peut regretter qu'il passe sous complet silence les travaux de l'école aixoise de recherche en syntaxe sur le français parlé. Ce chapitre est néanmoins moins solidement arrimé à un canevas théorique sous-jacent que celui concernant la tradition grammaticale tchèque. La tradition francophone de recherches sur la catégorie des particules donne l'impression d'être bien plus hétérogène, ce qu'on pourrait sans doute mettre sur le compte des conditions d'expression lexicale des équivalents de ce qu'on appelle « particules » dans les langues slaves, puisqu'en français rien n'empêche de considérer comme particules aussi bien des formes comme « voyez-vous » que « justement », « décidément », « à vrai dire », « éh bien », dont les procédés de construction sont on ne peut plus composites. L'auteur signale à quel point les langues germaniques et slaves sont plus riches en particules que le français, voire les langues romanes, mais on pourrait également ajouter, plus transparentes en termes de procédés de formation de ces unités lexicales et fonctionnelles.

Après cet itinéraire terminologique à cheval entre deux traditions de desciption grammaticale, l'auteur nous fait entrer dans le noyau dur de sa recherche : la partie empirique, conçue comme une étude de cas, sur deux particules du tchèque moderne : ***vždyt'*** et ***přece***. Le choix de ces deux paradigmes est d'autant plus justifié que ces deux « particules » se situent à la limite entre le plan conjonctif, en morphosyntaxe, et le plan discursif, en grammaire d'énonciation. Leur fonctionnement en tchèque rappelle celui d'un adverbe aussi central que *pure* dans la grammaire de l'italien, pour prendre un terme de comparaison dans une autre langue romane que le français, qui manque d'équivalents directs en traduction, ce qui rend l'enjeu de l'étude contrastive de Martin Svášek d'autant plus stratégique pour le champ de la traduction automatique. Le chapitre 4, quoique intitulé « étude empirique parallèle » remplit moins les promesses de parallélisme entre valeur fonctionnelle et sémantique de ces deux items en tchèque et en français qu'elle ne donne une description étymologique de ces deux unités, si bien qu'on peut considérer que l'auteur adopte une tactique d'approche des deux objets par leur motivation et par la diachronie – de manière somme toute très classique, voire préstructuraliste. Les figures 18 à 20 tirées de Diakorp demanderaient à être davantage explicitées, mais suffisent à montrer que ces deux items se

succèdent dans l'histoire de la langue, *přece* (*<před se(be))*étant plus anciennement implanté dans la langue que *vždyť* (< *veš-dy-ť*).

C'est au chapitre 5 qu'on aborde l'analyse des particules ***vždyť*** et ***přece*** à travers le corpus Fratchèque de Martin Svášek, tout en testant sa validité à l'aide d'autres grands corpora (5.1, 5.2.1-3, 5.3), comme SYN2000, à dominante journalistique (60%) et spécialisée (25%), et SYN2005, réparti entre 40% de littérature et près d'un tiers de textes littéraires et journalistiques. L'auteur démontre, par la congruence des profils statistiques de ces deux particules entre son corpus et d'autres corpora comme SYN, comparés au comportement du démonstratif ***to*** qui lui sert de *tertium comparationis* (en 5.3.), à la fois la robustesse de son corpus et la puissance de ses algorithmes (analyse A, p. 103 et analyse B, p. 116) et de ses scripts de détection de genres discursifs lui servant de cadre d'identification des occurrences énonciatives (et non pas conjonctives) de ***vždyť*** et ***přece***. On suit avec un vif intérêt la démarche investigatrice de l'auteur, qui traque ses données et les cadres opérationnels pour l'analyse des items – notamment le discours direct -, en déjouant les pièges liés aux erreurs de lecture optique énumérés dans la partie 2, ou partie B de la thèse. Martin Svášek conçoit et utilise les scripts comme autant d'outils pour tester ses hypothèses de catégorisation entre le plan discursif et le plan grammatical ou lexical dans la masse de ses textes parallèles. Dans sa traque des valeurs énonciatives et pragmatiques de ses deux objets ***vždyť*** et ***přece***, Martin Svášek procède avec l'habileté du trappeur : après avoir présenté les conditions de formation étymologique, donné des équivalents français et pondéré les occurrences dans divers corpora, il les fait passer par un filtre de périphrases de substitution, telles que *« C'est une surprise parce que »..., « Il faut penser au fait que... », « Il va de soi que... », « Puisque... »* dans la section 5.7. (intitulée modestement «vers une description lexicographique », alors que la démarche mise en œuvre n'a rien de lexicaliste, et relève plutôt d'une approche pragmatico-énonciative sobre, mais efficace) , abondamment illustrée par des tableaux à trois colonnes mettant en parallèle différents extraits de corpus représentatifs (pp. 125-175). Il adopte une grille de deux grandes classes de *métaphrasèmes* relevant, du point de vue pragmatique, de l'expression de la surprise (classe 1. : *« C'est une surprise parce que »*...) et de la mise en évidence (classes 2.a : *« Il faut penser au fait que... »*, 2b. *« Il va de soi que... »*, 2.c *« Puisque... »*) qui correspondent implicitement à des modalités exclamatives et assertives, mais filtrées par la prise en charge de l'énoncé par l'énonciateur. Ce choix méthodologique est excellent, car il permet de montrer à quel point l'analyse pragmatico-discursive est différente de la simple traduction, une particule comme ***vždyť*** par exemple pouvant aussi bien se traduire pour la classe 1 de l'expression de la surprise ou de l'incrédulité par *« mais c'est... »* ou *« pourtant »*, *« mais enfin »*, voire – de l'avis du présent rapporteur – à l'oral ou à l'oral littéralisé par des locutions comme *« ça alors... »*, *« éh bien mon vieux... »*, *« bon sang !»*. Or, là n'est pas la question, en termes de grammaire entendue comme système de représentations plutôt que comme système d'équivalences lexicales ou locutives. Le schéma de la figure 35 page 144 montre combien ces classes et sous-classes se recoupent partiellement.

A partir de 5.7.2, p. 155 et jusqu'à la conclusion l'auteur développe une fascinante étude des collocations des variantes de ses deux items particulaires, dûment pondérée dans son corpus. Le tableau 31, p. 156 illustre à merveille le procédé de la composition particulaire, ou séquences de particules composées, avec ***přece*** pour tête, et des particules ou surtout des conjonctions adjointes sur deux positions senestres et deux positions destres. Comme beaucoup de tableaux présentés dans cette thèse, une trame de graphe se dessine en filigrane, faisant non seulement apparaître toutes les combinaisons attestées, comme *ale **přece** jen*, *avšak **přece** jen*, ***přece** jen však*, etc., mais également leur poids statistique dans le corpus. Le présent rapporteur voit le très grand intérêt d'une telle démarche dans la description de pans entiers du fonctionnement de diverses langues du monde dont les grammaires descriptives

sont encombrées de gloses grammaticales baroques, faute d'avoir su voir les mécanismes de combinatoire conjonctive et particulaire en cause. En 5.8., l'auteur montre en quoi ***přece*** peut également jouer un rôle de focus exclamatif, et met en regard les équivalents traduits en français de ***přece*** avec les occurrences dans les textes en tchèque. L'effet de zone grise pour le traducteur du statut à la fois conjonctif, modal et énonciatif des deux particules se fait fortement sentir, d'une manière que l'auteur parvient à quantifier avec précision.

Enfin, dans sa conclusion, l'auteur reconnaît les limites de sa démarche, notamment quant aux rapports qu'entretiennent certains types de particules avec la structure communicative des énoncés, mais il pense avoir apporté des outils robustes et puissants pour automatiser des procédures de recherche allant dans ce sens pour de grands corpus, comptant des centaines de millions de mots. C'est en effet tout l'intérêt du T.A.L., qui représente pour la linguistique moderne une avancée au moins aussi décisive que le fut en son temps la naissance du comparatisme au début du XIXe siècle, ou l'essor du structuralisme européen autour des écoles de Prague et de Copenhague au milieu du XXe siècle. Par le T.A.L., les sciences du langage à l'orée du XXIe siècle se trouvent dotées d'outils d'observation, de classification et d'élucidation sur une échelle de grandeur inconnue auparavant. Cette thèse contribue à le démontrer et à le mettre en pratique.

En conclusion, cette thèse est, dans les conditions rédactionnelles exigées par l'exercice de cette co-tutelle, une excellente thèse d'ingénierie linguistique, mais aussi de théorie grammaticale assistée par le T.A.L. Martin Svášek y démontre d'indéniables qualités de programmateur, de chercheur, de traducteur et de rédacteur. La recherche menée à bien par Martin Svášek dans ce travail est solidement ancrée dans la linguistique quantitative et la linguistique générale du XXIe siècle. Elle contribue à faire avancer de manière originale la méthodologie en typologie linguistique et en linguistique contrastive romano-slave, ce qui n'est pas peu dire. Elle mérite d'être soutenue publiquement dans les plus brefs délais. Nul doute qu'elle donnera lieu à des débats féconds sur le rôle du T.A.L. dans le renouveau des Sciences du Langage en Europe.

Jean Léo Léonard, Maître de Conférences habilité à diriger des recherches, Université de Paris 3 et UMR 70 18 CNRS.

Fait à Paris, le 10/11/2007

J. Léonard