

**Univerzita Karlova
v Praze**

Filozofická fakulta

Ústav teoretické
a počítačové lingvistiky
(ÚTKL)

filologie – matematická
lingvistika

**Institut National des
Langues et Civilisations
Orientales à Paris**

Langues, Logiques,
Informatique et Cognition –
Centre d'Etudes et de
recherche en TAL
(LALIC – CERTAL)

Études tchèques – option
Traitement automatique des langues

Martin S v á š e k

**Définition, élaboration et exploitation d'un
corpus parallèle bidirectionnel
français – tchèque tchèque français**

**Návrh, vytvoření a využití francouzsko-
českého a česko-francouzského
paralelního korpusu**

**Parallel bi-directional Czech-French and
French-Czech corpus : definition, design,
build-up and exploitation**

Teze disertační práce

Résumé de thèse

vedoucí práce

directeur de thèse

Doc. RNDr. Vladimír Petkevič,
CSc.

Prof. PhDr. Patrice Pognan, DrSc.

Fratchèque

Un corpus parallèle bidirectionnel français-tchèque tchèque-français : définition, élaboration et exploitation

La thèse est exposée en trois parties en accord avec son titre. Dans un premier temps, nous introduisons le concept de corpus parallèle et, sur le plan général, le définissons comme un ensemble de textes en plusieurs langues (au moins deux), constitué de couples original-traduction. Il faut le distinguer des corpus comparables et des mémoires de traduction qui sont construits avec un autre objectif et qui fonctionnent d'une façon complètement différente. Ainsi, une terminologie pour nommer les différents types d'ensembles de textes en plusieurs langues est mentionnée. Pour avoir une vue générale du domaine en question, les lecteurs peuvent prendre connaissance des corpus parallèles existants à l'heure actuelle. Le projet d'un corpus parallèle bidirectionnel français-tchèque tchèque-français (*Fratchèque*) est ensuite défini et cela en vue de son usage futur pour les recherches linguistiques, notamment pour la recherche des expressions non fléchies qui représentent le centre d'intérêt de l'auteur. *Fratchèque* est un corpus parallèle de ressources écrites dont les textes en français et en tchèque proviennent de la littérature. Son volume est comparable à celui de la plupart des corpus parallèles mentionnés précédemment car il correspond à un million de mots pour chaque langue. Comme il s'agit d'un travail individuel (à la différence de la majeure partie des projets comparables), nous énumérons les difficultés qui ont empêché d'inclure dans *Fratchèque* également d'autres types de texte. Elles résultent notamment du manque de ressources bilingues tchèque/français de qualité.

Le corpus existe uniquement sous forme numérique afin de permettre les recherches à l'aide d'un ordinateur. Il prétend représenter la langue moderne, si bien que seuls les textes provenant de la période postérieure à l'année 1945 sont pris en compte. La structure des fichiers stockés sur un disque dur et gérés par le logiciel ParaConc pourrait être enrichie dans l'avenir de nouveaux couples de textes tchèque/français ou français/tchèque. Elle n'est pas marquée explicitement par des balises XML ce qui est justifié par la non nécessité actuelle du balisage (le logiciel utilisé pour le traitement du corpus n'en a pas besoin). Cette tâche, considérée comme importante, est reportée à plus tard. Seules les fins de phrase et les fins de paragraphe sont balisées à l'aide d'un macro de Word ce qui a rendu plus précise la phase de l'alignement.

Tenant compte des limitations existantes liées à la recherche des textes susceptibles d'être incorporés dans *Fratchèque*, une liste de critères de choix, présentés par ordre d'importance, est établie. L'impossibilité de constituer un corpus représentatif (à base des critères sociolinguistiques sophistiqués et d'une taille qui correspondrait à quelques dizaines de millions de mots) par une seule personne est compensée par une tentative de respecter le principe de la plus grande diversité du choix possible. Ainsi, l'influence du style d'un auteur ou d'un traducteur sur différents aspects de la langue représentée par le corpus devrait être

réduite au minimum. Il faut donc éviter l'inclusion dans le corpus de plusieurs titres d'un même auteur ou traducteur. Le tableau 1 résume tous les critères respectés avec un poids relatif qui leur a été attribué pour indiquer leur importance :

CRITÈRES DE CHOIX POUR SÉLECTIONNER LES TEXTES À INCLURE DANS FRATCHÈQUE

Entité concernée	Critère	Poids relatif	Critère n°
Œuvre/traduction	Date de parution de l'original	25	(1)
	Date de parution de la traduction	25	(2)
	Disponibilité du titre	10	(3)
	Nombre de dialogues	10	(4)
	Correspondance d'éditions	5	(5)
	Qualité matérielle de l'édition ¹	5	(6)
Auteur/traducteur	<i>La plus grande diversité d'auteurs et de traducteurs possible</i>	10	(7)
	Réputation du traducteur	10	(8)
		100	

Tableau 1

En dépit des difficultés liées au respect des droits d'auteur, le choix des textes s'est orienté vers les textes littéraires car cela devrait éviter avant tout les hésitations sur la qualité de la production écrite disponible sur Internet et libre de droits. D'ailleurs, la recherche de textes bilingues sur Internet présenterait un facteur trop limitatif lors d'une telle entreprise. Essayant de garantir la modernité des langues en question, représentées par le corpus, nous avons constitué *Fratchèque* des titres cités dans le tableau 2.

¹ Pour la reconnaissance optique de caractères, il est préférable de scanner les textes sans défauts d'impression apparents (le contraste faible des caractères imprimés par rapport à la couleur du papier qui peut produire un effet de « ponctuation transparente » où on ne peut plus distinguer les points fins, par exemple).

ORIGINAL			TRADUCTION		
Auteur	Titre	1 ^{ère} éd.	Titre	Traducteur	1 ^{ère} éd.
Assouline, Pierre	Double vie	2000	Dvojí život	Šárka Belisová	2004
Assouline, Pierre	La cliente	1998	Zákaznice	Lubomír Martínek	2000
Beigbeder, Frédéric	99 francs	2000	99 franků	Markéta Demlová	2003
Laurens, Camille	L'amour, roman	2003	Láska, román	Alexandra Pfimplová	2004
Mandiargues, A. P. de	Soleil des loups	1951	Vičí slunce	Ladislav Šerý	1992
Page, Martin	Comment je suis devenu stupide	2001	Jak jsem se stal hlupákem	Alan Beguivin	2003
Sarn, Amélie	Elle ne pleure pas, elle chante	2002	Nepláče, zpívá	Andrea Prošková	2004
Tournier, Michel	Le Coq de bruyère	1978	Tetřev hlušec	Václav Jamek	1984
Toussaint, Jean-Philippe	Faire l'amour	2002	Milovat se	Jovanka Šotolová	2004
Tristan, Frédéric	Les tribulations héroïques de B. Kober	1980	Hrdinné útrapy B. Kobera	Oldřich Kalfiřt	2003
Fuks, Ladislav	Spalovač mrtvol	1967	L'incinérateur de cadavres	Barthélémy Müller	2004
Hrabal, Bohumil	Příliš hlučná samota	1980	Une trop bruyante solitude	Max Keller	1983
Hrabal, Bohumil	Postřižiny	1976	La chevelure sacrifiée	Claudia Ancelot	1987
Kundera, Milan	Žert	1967	La plaisanterie	Marcel Aymonin	1968
Kundera, Milan	Žert		La plaisanterie	Claude Courtot et l'auteur	1980
Ouředník, Patrik	Europeana	2001	Europeana	Marianne Canavaggio	2004
Pecka, Karel	Pasáž	1976	Passage	Barbora Faure	1990
Škvorecký, Josef	Zbabělci	1958	Les lâches	Françoise London-Daix	1977
Trefulka, Jan	Veliká stavba	1973	Le grand chantier	Barbora Faure	1999
Třešňák, Vlastimil	U jídla se nemluví	1996	On ne parle pas la bouche pleine	Marianne Canavaggio	2000
Weil, Jiří	Na střeše je Mendelssohn	1965	Mendelssohn est sur le toit	Erika Abrams	1993

Tableau 2

La partie linguistique de la thèse s'appuie sur la réalisation préalable du corpus parallèle. Elle aborde un phénomène (une catégorie lexicale ?) connu en tchèque sous le terme *částice* qui n'a pas d'équivalent univoque en français. L'étude commence donc par un panorama terminologique à travers les langues – à la recherche d'une étiquette universelle existante pour un phénomène abordé différemment dans les différentes traditions linguistiques. Tandis que l'approche tchèque correspond par de nombreux aspects au point de vue allemand et aborde une catégorie connue dans de nombreuses traditions slaves (p. ex. russe, bulgare), la question n'est pas traitée de la même façon dans le cas de l'anglais et du français. Le parcours terminologique s'arrête aux *mots du discours* et *particules énonciatives* – termes le plus souvent liés en français à la question. Le premier est attribué à Oswald Ducrot et apparaît dans sa théorie de l'argumentation, l'autre a été choisi comme titre de la monographie de Jocelyne Fernandez (1994) mais apparaît déjà dans les travaux des slavistes francophones dans les années 1980.

L'élaboration du corpus est décrite d'une façon détaillée en suivant de près toutes les démarches et tout le paramétrage des logiciels utilisés qui sont nécessaires pour transformer

les livres sélectionnés en corpus parallèle. Le travail commence avec le logiciel de reconnaissance optique de caractères FineReader. Celui-ci réussit la numérisation de textes avec un taux de fiabilité assez élevé, pourtant, le résultat demande une vérification manuelle de nombreux passages. Nous contrôlons la qualité et l'encodage des textes numérisés sous MS Word 2002. Ce n'est qu'ensuite qu'on procède à la constitution d'un corpus parallèle géré par ParaConc. L'alignement de textes au niveau des paragraphes et des phrases étant une tâche cruciale pour le fonctionnement correct du corpus, le côté théorique aussi bien que le côté pratique de l'alignement sont abordés. La description des difficultés rencontrées lors de l'élaboration du corpus met en évidence les défauts des différents logiciels utilisés. Elle reprend d'abord les erreurs les plus fréquentes de FineReader qui sont organisées selon leur typologie, ensuite le manque de flexibilité de ParaConc lors de l'alignement. Nous proposons également des solutions pour résoudre les problèmes identifiés (il s'agit notamment du problème avec les points de suspension et les particularités liées à l'encodage lors de l'exportation des fichiers texte en dehors du système). Nous avons décidé de remettre le projet d'étiquetage morphologique à plus tard car la complexité de cette tâche ralentirait fortement le travail – sans pour autant fournir une garantie que le balisage ainsi effectué ait un impact positif considérable sur le résultat de l'étude consacrée aux mots non fléchis.

La tradition tchèque connaît le terme *částice* depuis au moins les années 1940 mais nous pouvons voir qu'il ne correspondait pas toujours à la même approche. Trávníček (1949) semble être le premier à l'employer pour désigner le phénomène en question. Les conceptions des autres grammairiens ont évolué dans le temps en suivant des approches soit convergentes (Havránek – Jedlička 1960, 1970, 1981; Mistrík 1959), soit complètement différentes (Kopečný 1958, Komárek 1978, Daneš 1985).

Selon Havránek – Jedlička (1960, p. 272), « [les particules sont] des mots non fléchis qui introduisent des énoncés indépendants et indiquent leur nature (ordre, question, exclamation) ou les différentes variétés modales ou expressives (*at', necht', kéž, což, copak* etc.). Elles-mêmes n'expriment pas de fonction syntaxique. Parce qu'elles ne lient pas les énoncés, il faut les différencier des conjonctions. » Dans Havránek – Jedlička (1960, p. 272), les auteurs mettent les *částice* en relation avec *větné členy* (éléments de la syntaxe d'une proposition, fonctions syntaxiques des mots autrement dit) : « Parce qu'elles ne lient pas les énoncés (ou les mots qui expriment les fonctions syntaxiques) il faut les différencier des conjonctions au sens propre. » Les différentes modifications témoignent d'une évolution chez Havránek – Jedlička vers le concept des mots au rôle 'serviteur' dans le cadre de la phrase (« slova se služebnou funkcí ve větě »). František Kopečný (1958, p. 143) voit dans *les částice* uniquement des « restes » de mots qui sont devenus morphèmes dépendants. Il est le seul à donner au terme *částice* (*partikule*) une définition qui ne les considère pas comme des mots à sens plein mais seulement comme des formants (*formanty*). Mistrík, linguiste slovaque, prône pour la recherche des critères positifs de délimitation des *částice* pour ne pas en faire « un réservoir de mots qu'on ne réussit pas à classer ». Mistrík (1959, p. 209) formule donc la définition suivante : « Les particules sont des mots non fléchis qui expriment la relation du

locutaire envers le choix des différentes expressions ou envers tout l'énoncé. » Dans Mistrík 1959 il construit une typologie de cette catégorie et présente un dictionnaire des *částice*. Dans Komárek (1978, p. 10), les *částice* sont délimitées comme la classe des mots a) qui ne font pas partie de la structure syntaxique d'une proposition ; b) qui ne sont pas éléments de la structure grammaticale de la phrase et c) qui ne sont pas un énoncé (« *větný člen – gramatický komponent věty – výpověď* »). Ces trois traits négatifs montrent comment distinguer les autres catégories lexicales de mots non fléchis :

	adverbes	conjonctions	onomatopées	particules
větný člen	+	+/-	-	-
gramatický komponent věty	+	+	-	-
výpověď	+	-	+	-

Tableau 3

František Daneš utilise dans ses travaux un cadre théorique particulier, son approche fonctionnelle ne distingue pas les différentes catégories de mots au sens traditionnel du terme. Il travaille avec les unités lexicales en partant de la fonction de ces dernières. Ainsi, nous trouvons dans Daneš (1985) les restricteurs (*restriktory*), contextualiseurs (*kontextualizátory*) *také, zase, přece* et les connecteurs (*konektory*) *stejně, jinak, alespoň*. Les termes renvoient à la fonction de ces expressions. Daneš ajoute une note terminologique qui fait le lien explicite avec la catégorie des *částice* du travail de Mistrík (1966). Selon Daneš (inspiré par Mistrík), les connecteurs ont une double fonction : en plus de celle des contextualiseurs (insérer les phrases dans un contexte), ils modifient en même temps le sens de l'énoncé.

Petit à petit, on découvre que les mots tchèques rangés dans cette catégorie jouent un rôle important sur plusieurs plans de description linguistique. Surtout les dernières grammaires (*Mluvnice češtiny 2, Mluvnice češtiny 3, Čeština – řeč a jazyk, Příruční mluvnice češtiny*) tentent de présenter leur fonctionnement dans sa complexité. Pour montrer celle-ci, il suffit de rappeler que les *částice* sont liées au plan prosodique (Mistrík 1959, Komárek 1978, *Příruční mluvnice češtiny*), à la structure communicative (Daneš) et jusqu'à la cohérence textuelle (*Mluvnice češtiny 3, Příruční mluvnice češtiny*). Sans les mentionner explicitement, les descriptions font souvent appel aux critères pragmatiques.

Un regard sur les ouvrages francophones confirme qu'il est presque impossible de trouver en français un groupe correspondant de moyens linguistiques relativement cohérent. Le terme *particules* n'est pas consistant, son contenu reste en français très vague. Seule la *Grammaire méthodique du français* (Riegel et alii 1998, p. 380) mentionne sous le titre « La marque d'une propriété globale de la phrase » certaines expressions à fonction similaire à celle des *částice*. D'autres exemples se trouvent dans une monographie consacrée à l'articulation du discours (Roulet et alii 1987) ainsi que chez Jocelyne Fernandez (1994). Ducrot (1980) étudie l'orientation argumentative de certains « connecteurs ». À la lumière de tous ces éléments, une approche purement empirique est adoptée, en réduisant le vaste domaine des *částice*

tchèques à deux exemples et leurs variantes. L'analyse commence donc par une note étymologique sur *přece* et *vždyť*, formes qui sont étudiées par la suite dans une perspective synchronique.

Les descriptions existantes insinuent qu'il y a une relation étroite entre ces mots et le discours et que ces mots devraient être fréquents surtout à l'oral. La première partie de l'étude tente alors de démontrer sur le corpus qu'il en est ainsi, tout en considérant que l'oral est repris à l'écrit le plus souvent entre guillemets. Ceci représente le point de départ pour l'analyse des occurrences de *vždyť*, *přece* et de leurs variantes *dyť*, *dyk*, *přec* et *přeci* dans le discours direct. Étant donné qu'il s'agit d'une propriété qui devrait se rencontrer dans les textes tchèques, la vérification se fait en deux étapes : d'abord sur les grands corpus de tchèque gérés par Ústav Českého národního korpusu² (analyse A), ensuite sur Fratchèque (analyse B). L'analyse A peut travailler avec les volumes de données significatifs d'ordre de centaines de millions de mots, l'analyse B se fait sur Fratchèque qui contient un million de mots environ pour chaque langue. Les deux sont réalisées d'une façon automatique – par des scripts rédigés dans le langage de programmation Python. L'analyse A doit se faire sur les données qui malheureusement ne respectent pas l'usage typographique tchèque des guillemets et elle est conditionnée par certaines restrictions qu'impose le logiciel pour la recherche dans les grands corpus ; l'algorithme de l'analyse A en tient compte. En revanche, l'analyse B est le fruit de l'application d'un autre script sur les données de Fratchèque non déformées et parfaitement accessibles. Pour obtenir un résultat irréprochable, les mesures effectuées sur le Corpus National Tchèque sont confrontées avec le même type d'analyse, centrée cette fois-ci sur un mot (le pronom *to*) qui ne devrait pas avoir de relation particulière avec le discours direct. Les mots *přece*, *přeci*, *vždyť*, *dyť* se trouvent dans le discours direct en moyenne de 1,7 jusqu'à 6,5 fois plus souvent que *to* ce qui confirme l'hypothèse initiale que les *částice* étudiées sont plus fréquentes à l'oral qu'à l'écrit. L'analyse B montre que *vždyť/dyť* se trouvent dans le discours direct dans plus de 60 pour cent des cas, le même chiffre pour *přece/přec* est presque de 40 pour cent. Les recherches dans tous les corpus accessibles montrent qu'entre les guillemets et l'occurrence des mots étudiés, il existe une relation de proportionnalité directe ce qui peut être interprété comme une preuve supplémentaire de l'hypothèse de départ.

L'étude continue avec l'analyse qualitative des données de Fratchèque afin de présenter systématiquement tous les types d'usage de *vždyť*, *přece* et de leurs variantes dans le but de proposer une description lexicographique pour un dictionnaire bilingue tchèque-français. Il convient de distinguer à cette fin d'un côté les collocations fréquentes, d'autre côté les énoncés où ces formes apparaissent seules. Dans ce dernier cas, quatre paraphrases et un test complémentaire servent à distinguer les différentes significations très peu nuancées les unes des autres. Les paraphrases semblent couvrir la plupart des occurrences testées mais certains chevauchements et certains doubles, voire triples remplacements possibles montrent que les

significations de *vždyt'* et *přece* sont réparties d'une façon continue sur une échelle. Seules celles qui se trouvent aux extrémités de cette échelle sont incompatibles.

Après cette classification, il est possible de regrouper de la même façon les traductions en français et d'en dégager les conclusions. Les données de Fratchèque prouvent que les expressions composées sont plus facilement traduites par un équivalent lexical tandis que les *částice* individuelles restent le plus souvent sans équivalent. Les exemples type sont repris dans les entrées lexicographiques qui résument toute l'analyse. Quelques exercices destinés aux apprenants du tchèque ou du français montrent comment utiliser le corpus bilingue dans la didactique des langues. Ils sont basés sur les résultats de l'étude linguistique. Juste avant de conclure, on discute quelques questions qui concernent la possibilité d'évaluer automatiquement la qualité de traductions liées à la présence de *částice*.

La conclusion mène une évaluation de la réalisation des objectifs établis, fait le point sur la problématique linguistique, indique les directions possibles pour la poursuite éventuelle de la recherche et énumère de nombreux types d'utilisation possibles de Fratchèque.

Bibliographie

- Altmann, H. (1976): *Die Gradpartikeln im Deutschen. Untersuchungen zu ihrer Syntax, Semantik und Pragmatik*. Niemeyer, Tübingen
- Asher, R. E., Simpson, J. M. Y. (éds) (1994): *The Encyclopedia of Language and Linguistics*. Oxford – New York – Seoul – Tokyo, Pergamon Press.
- Anscombe, J.-C. (2002): "*Mais/pourtant* dans la contre-argumentation directe: raisonnement, généralité, et lexique", *Linx*, 46, pp. 115–131.
- Aslanides, S. (2001): *Grammaire du français. Du mot au texte*. Paris, Honoré Champion Éditeur.
- Barlow, M. (1995): "ParaConc – A Concordancer for Parallel Texts", *Computers and Texts*, 10.
- Barnes, B. K. (1995): "Discourse Particles in French Conversation : (*eh*) *ben*, *bon*, and *enfin*", *The French Review*, 68, (5), pp. 813–822.
- Bauer, J. (1964): "Spojky a částice", *Sborník prací filosofické fakulty Brněnské university*, 13, pp. 131–139.
- Bauer, J., Grepl, M. (1980³): *Skladba spisovné češtiny*. Praha, SPN.
- Bělič, J. – Kamiš, A. – Kučera, K. (1979): *Malý staročeský slovník*. Praha, SPN.
- Běličová, H. (1987): "M. Grochowski: Polskie partykuly. Składnia, semantyka, leksykografija. Wrocław – Warszawa – Kraków – Gdańsk – Łódź, Ossolineum, 1986", *Slavia*, 56, pp. 285–290.
- Biber, D., Conrad, S. et Reppen, R. (1998): *Corpus Linguistics : Investigating Language Structure and Language Use*. Cambridge, Cambridge University Press.

- Bilger, M. (éd.) (2000): *Corpus: Méthodologie et applications linguistiques*. Paris – Perpignan, Honoré Champion Éditeur et les Presses Universitaires de Perpignan.
- Borin, L. (éd.) (2002): *Parallel corpora, parallel worlds*. Amsterdam – New York, Rodopi.
- Botley, S. P., McEnery, A. M. et Wilson, A. (éds.) (2000): *Multilingual Corpora in Teaching and Research*. Amsterdam – Atlanta, Rodopi.
- Bronckart, J.-P., Schneuwly, B. (1984): "La production des organisateurs textuels chez l'enfant", Moscato, M., Piérault-Le-Bonniec, G. (éds), *Le langage. Construction et actualisation*. Rouen, Presses Universitaires de Rouen, pp. 165–178.
- Brown, P. F., Lai, J. et Mercer, R. (1991): "Aligning Sentences in Paralel Corpora", *Proceedings of ACL-91*.
- Cervoni, J. (1987): *L'énonciation*. Paris, PUF, coll. Linguistique nouvelle.
- Čechová, M. et alii (2000): *Čeština – řeč a jazyk*. 2^e édition, corr. Praha, ISV
- Čermák, F. et alii (1988): *Slovník české frazeologie a idiomatiky. Výrazy neslovesné*. Praha, Academia
- Čermák, F., Klímová, J. et Petkevič, V. (éds) (2000): "Studie z korpusové lingvistiky", *AUC Philologica*, 3-4/1997. Praha, Nakladatelství Karolinum.
- Čermák, F., Klégr, A. (2004): *Modality in Czech and English. Possibility particles and the conditional mood in a parallel corpus*. Amsterdam – Philadelphia, John Benjamins Publishing Company.
- Červená, V., Churavý, M., Machač, J., Mejstřík, V. et alii (éds.) (1988): *Slovník české frazeologie a idiomatiky. Výrazy neslovesné*. Praha, Academia.
- Dalmas, M. (1996): "Particules et autres mots du discours", *Nouveaux Cahiers d'allemand*, 14 (4), pp. 437–441.
- Daneš, F. (1957): *Intonace a věta ve spisovné češtině*. Praha, Academia.
- Daneš, F. (1974): "Functional sentence perspective and organization of the text", *Papers on functional sentence perspective*, pp. 100–128.
- Daneš, F. (1985): *Věta a text*. Praha, Academia.
- Daneš, F., Hlavsa, Z. et alii (1987): *Větné vzorce v češtině*. Praha, Academia.
- Daniellsson, P. (2005): "Constitution of Corpus... ", transparentes pour le séminaire tenu à Brighton du 15 au 17 mars, accessibles sur l'adresse internet suivante : <http://www.humcorp.bham.ac.uk/Presentations/Parallel%20Texts%20-%20AHRB.ppt#2> (accès vérifié en octobre 2007)
- Deloffre, F. (1955): *Marivaux et le marivaudage*. Paris, Les Belles Lettres.
- Denis, D., Sancier-Chateau, A. (1997): *Grammaire du français*. Édition Encyclopédie d'aujourd'hui, La Pochothèque, Paris, Librairie générale française.
- Dubois, J. – Lagane, R. (1973): *La nouvelle grammaire du français*. Paris, Larousse.
- Ducrot, O. et alii (1980): *Les mots du discours*. Paris, Minuit.
- Ducrot, O. (1988): "L'argumentation dans la langue : bibliographie", *Modèles linguistiques*, 10 (2) ("Analyse transphrastique"), pp. 131–132.

- Duden Grammatik* (1984): *Grammatik der deutschen Gegenwartssprache*. 4e édition. Bibliographisches Institut, Mannheim
- Edmondson, W., House, J. (1981): *Let's talk and talk about it*. Munich, U & S Pädagogik.
- Eisenberg, P. (1989): *Grundriß der deutschen Grammatik*, 2^e édition. Stuttgart, Metzler.
- Fernandez Bravo, N. (1999): "La traduction des particules de discours et des particules de focalisation dans le sens allemand/français comme révélateur de l'interprétation textuelle", *Nouveaux Cahiers d'allemand*, 17 (3), pp. 595–605.
- Filipec, J., Daneš, F. (éds.) (1978): *Slovník spisovné češtiny pro školu a veřejnost*. Praha, Academia.
- Gale, W., Church, K. W. (1991): "A Program for Aligning Sentences in Bilingual Corpora", *Proceedings of ACL-91*.
- Gale, W., Church, K. W. (1993): "A Program for Aligning Sentences in Bilingual Corpora", *Computational linguistics*, 19 (1), pp. 75–102.
- Garvin, P., Brewer, J. et Mathiot, M. (1967): "Predication-Typing. A pilot study in semantic analysis", *Language Monograph*, 27, Washington.
- Grepl, M. (1989): "Partikulizace v češtině", *Jazykovědné aktuality*, 26, pp. 95–100.
- Grepl, M. (1979): "Úvodní poznámky k tzv. jistotní modalitě". *SaS*, 40, pp. 81–87.
- Grepl, M., Karlík, P. (1989): *Skladba spisovné češtiny*. Praha, SPN.
- Grepl, M. et alii (1995): *Příruční mluvnice češtiny*. Praha, NLN.
- Grevisse, M., Goose, A. (1998): *Le bon usage*. 14^e éd., Bruxelles, Duculot.
- Grochowski, M. (1986): *Polskie partykuly. Składnia, semantyka, leksykografija*. Wrocław – Warszawa – Kraków – Gdańsk – Łódź, Ossolineum.
- Grochowski, M. (1979): "Założenia opisu polskich partykuł", *Slavica Ludensia*, 7, pp. 93–100.
- Groupe λ-1 (1975): "Car, parce que, puisque", *Revue Romane* 10, 248–280.
- Grünthal-Robert, T. (1997): *Particules énonciatives en français, en finnois et en estonien*. Thèse de doctorat, Université Paris III.
- Gülich, E. (1970): *Makrosyntax der Gliederungssignale im gesprochenen Französisch*. München, Fincke.
- Habert, B., Fabre, C. et Isaac, F. (1998): *De l'écrit au numérique*. Paris, Masson.
- Havránek, B., Jedlička, A. (1960): *Česká mluvnice*. Praha, SPN.
- Havránek, B., Jedlička, A. (1970): *Česká mluvnice*. 2^e éd., Praha, SPN.
- Havránek, B., Jedlička, A. (1981): *Česká mluvnice*. 3^e éd., Praha, SPN.
- Havránek, B. et alii (éds.) (1971): *Slovník spisovného jazyka českého*. 4 tomes. Praha, Academia.
- Heidolph, K.-E., Flämig, W. et Motsch, W. (1981): *Grundzüge einer deutschen Grammatik*. Akademie-Verlag, Berlin

- Helbig, G., Buscha, J. (1988): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. 11^e édition. Verlag Enzyklopädie, Leipzig
- Helbig, G., Kötz, W. (1981): *Die Partikeln*. Verlag Enzyklopädie, Leipzig.
- Hentschel, E., Weydt, H. (1990): *Handbuch der deutschen Grammatik*. de Gruyter, Berlin
- Hoffmanová, J. (1983): *Sémantické a pragmatické aspekty koherence textu*, *Linguistica* VI, Praha.
- Jayez Annemasse, J. (1982): "Quand bien même *pourtant*, *pourtant quand même*", *Cahiers de linguistique française*, 4, pp. 189–217.
- Jocelyne Fernandez, M. M. (1994): *Les particules énonciatives*. Paris, PUF.
- Jungmann, J. (1835-1839): *Slownjk česko-německý Josefa Jungmanna*. Praha, Řjzenjm W. Spinky
- Karlík, P., Nekula, M. et Pleskalová, J. (éds.) (2002): *Encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny.
- Kay, M. et Röscheisen, M. (1993): "Text Translation Alignment ", *Computational Linguistics* 19 : 1, pp. 121-142.
- Komárek, M. (1958): "K sémantické a syntaktické charakteristice slov *ještě* a *již (už)*", *Studie ze slovanské jazykovědy*, pp. 147–152.
- Komárek, M. (1979): "K jednomu funkčnímu rozdílu v soustavě partikulí", *SaS*, 40, pp. 139–142.
- Kopečný (1962): *Základy české skladby*. 2^e éd., révisée. Praha, SPN.
- Kopečný, F. et alii (1973): *Etymologický slovník slovanských jazyků. Slova gramatická a zájmena*, tome 1, *Předložky, koncové partikule*. Praha, Academia.
- Kopečný, F. et alii (1980): *Etymologický slovník slovanských jazyků. Slova gramatická a zájmena*, tome 2, *Spojky, částice, zájmena a zájmenná adverbia*. Praha, Academia.
- Kouloughli, D. E. (1987): "Les particules ont-elles un sens ?", *BULAG*, 1986–1987 (13), pp. 215–237.
- Le Goffic, P. (1993): *Grammaire de la phrase française*. Paris, Hachette.
- Luzzati, D. (1982): "« Ben » appui du discours", *le Français moderne*, 50, pp. 193–208.
- Machek, V. (1968): *Etymologický slovník jazyka českého*, 2^e éd., révisée et complétée, Praha, Academia.
- Martinet, A. et alii (1979): *Grammaire fonctionnelle du français*. St. Cloud, CREDIF/Paris, Didier.
- Mason, O. (2000): *Programming for Corpus Linguistics*. Edinburgh, Edinburgh University Press.
- Melamed, I. D. (1999): "Bitext Maps and Alignment via Pattern Recognition", *Computational linguistics*, 25 (1), pp. 107–130.
- Mellet, S. (2002): "Corpus et recherches linguistiques", *Corpus*, 1, pp. 5–12.

- Metrich, R. (1985): "Proposition pour le traitement des « mots du discours » en lexicographie bilingue français-allemand: l'exemple de *d'accord, donc* et *enfin*", *Nouveaux Cahiers d'allemand*, pp. 33–52.
- Metrich, R. (1985): "Proposition pour le traitement des « mots du discours » en lexicographie bilingue français-allemand: l'exemple de *d'accord, donc* et *enfin* II: *donc*", *Nouveaux Cahiers d'allemand*, 3 (2), pp. 159–172.
- Metrich, R. (1985): "Proposition pour le traitement des « mots du discours » en lexicographie bilingue français-allemand: l'exemple de *d'accord, donc* et *enfin* III: *enfin*", *Nouveaux Cahiers d'allemand*, 2, pp. 237–247.
- Metrich, R. (1986): "Petit dictionnaire bilingue permanent des interjections, particules et autres « mots du discours »", *Nouveaux Cahiers d'allemand*, 4, pp. 367–386.
- Metrich, R. (1987): "Petit dictionnaire bilingue permanent des interjections, particules et autres « mots du discours » III : *bien*", *Nouveaux Cahiers d'allemand*, 5 (2), pp. 121–139.
- Meyer, Ch. F. (2002): *English Corpus Linguistics. An introduction*. Cambridge, Cambridge University Press.
- Mistrík, J. (1959): "K otázke častíc v slovenčine", *Jazykovedné štúdie 4. Spisovný jazyk*, pp. 201–228.
- Mistrík, J. (1974): "Častice a štýly", *Jazykovedné štúdie 12. Spisovný jazyk*, pp. 137–141.
- Nekula, M. (1996): *System der Partikeln im Deutschen und Tschechischen*. Tübingen, Niemayer.
- Němec, I. (1968): *Vývojové postupy české slovní zásoby*. Praha, Academia.
- Paillard, D. et alii (1986): *Les particules énonciatives en russe contemporain*. Paris, Institut d'études slaves.
- Panevová, J. (1980): *Formy a funkce ve stavbě české věty*. Praha, Academia.
- Peciar, Š. (1959): "K diskusii o časticiach, Slovenská reč XXII, 1957, 47–54", *Niektoré problémy klasifikácie neohybných slov, Sborník k 70. narodeninám akad. Fr. Trávníčka*, Praha 1958, 141
- Petr, J. et alii (1986): *Mluvnice češtiny (2). Tvarosloví*. Praha, Academia.
- Petr, J. et alii (1987): *Mluvnice češtiny (3). Skladba*. Praha, Academia.
- Porquier, R. (2001): "Mots-phrases, phrasillons, locutions-énoncés : aux frontières de la grammaire et du lexique en français langue étrangère", *Langue française*, 131, pp. 106–123.
- Příruční slovník jazyka českého*. (1953-1955). Praha, Academia.
- Rangelova, A. (1993): *Hodnotící částice v české větě v porovnání s bulharštinou*. Kandidátská disertace, Praha, ČSAV.
- Riegel, M., Pellat, J.-Ch. et Rioul, R. (1998): *Grammaire méthodique du français*. 4^e éd., Paris, PUF.
- Rosen, A. (2005) : "In search of the best method for sentence alignment in parallel texts" dans Garabík, R. (éd). *Computer Treatment of Slavic and East European Languages: Third International Seminar, Bratislava 10-12 November 2005*, Bratislava, VEDA, pp. 174-185.

- Roulet, E. et alii (1987): *L'articulation du discours en français contemporain*. 2^e éd. P. Lang, Paris/Bern.
- Salkie, R. (2000): "Quelques questions méthodologiques dans l'exploitation des corpus multilingues" dans Bilger, M. (éd). *Corpus: Méthodologie et applications linguistiques*. Paris– Perpignan, Honoré Champion Éditeur et les Presses Universitaires de Perpignan, pp. 180–195.
- Simard, M., Foster, G. et Isabelle, P. (1992): "Using cognates to Align Sentences in Bilingual Corpora", *Proceedings of TMI-92*.
- Slovník spisovné češtiny pro školu a veřejnost*. (2003). 3^e éd., corr. Praha, Academia.
- Slovník spisovného jazyka českého* (1989). 8 tomes. Praha, Academia.
- Šmilauer, V. (1966): *Novočeská skladba*. 2^e éd. Praha, SPN.
- Thurmair, M. (1989): *Modalpartikeln und ihre Kombinationen*. Tübingen, Niemeyer.
- Tognini-Bonelli, E. (éd). (2001): *Studies in Corpus Linguistics*. Amsterdam/Philadelphia, John Benjamins Publishing Company.
- Trávníček, F. (1948): *Mluvnice spisovné češtiny. Část I: Hláskosloví, tvoření slov, tvarosloví*. Praha, Melantrich.
- Trávníček, F. (1948): *Mluvnice spisovné češtiny. Část II: Skladba*. Praha, Melantrich.
- Trávníček F. (1951): *Mluvnice spisovné češtiny. Část II: Skladba*. Praha, Slovanské nakladatelství.
- Tricás, M. (1990): "L'argumentation concessive française et espagnole: le transfert en espagnol de quelques connecteurs (pourtant, quand même, même si, or...) et d'autres formules concessives", *Meta*, 35 (3), pp. 529–537.
- Uhlířová, L. (1979): "K postavení tzv. větných příslovcí v aktuálním členění", *SaS*, 40, pp.143–148.
- Vachek, J. (1972): "Glosa k novočeskému *vždyt*", *NŘ*, 55, pp. 94–97.
- Veiland, R. (1998): "*Quand même et tout de même*: concessivité, synonymie, évolution", *Revue Romane*, 33 (2), pp. 217–245.
- Véronis, J. (éd.) (2000): *Parallel Text Processing. Alignment and Use of Translation Corpora*. Dodrecht – Boston – London, Kluwer Academic Publishers.
- Véronis, J., Langlais, P. (1999). ARCADE: évaluation de systèmes d'alignement de textes multilingues. In K. Chibout, F. Néel, J. Mariani & N. Masson (Eds.), *Ressources et Evaluation en Ingénierie de la Langue* (pp. 77-100): Aupelf-Uref.
- Vlasák, V., Lyer, S. (1993): *Česko-francouzský slovník. I, II*. 2^e éd., Praha, SPN.
- Vondráček, M. (1998): "Citoslovce a částice – hranice slovního druhu", *NŘ*, 81 (1), pp. 29–37.
- Vondráček, M. (1999): "Příslovce a částice – hranice slovního druhu", *NŘ*, 82 (2), pp. 72–78.
- Wagner, R.-L., Pinchon, J. (1962): *Grammaire du français classique et moderne*. Paris, Hachette.
- Weinrich, H. (1989): *Grammaire textuelle du français*. Paris, Didier/Hatier.

- Wilmet, M. (1997): *Grammaire critique du français*. Paris – Bruxelles, Hachette/Duculot.
- Winther, A. (1985): "Bon (bien, très bien): ponctuation discursive et ponctuation métadiscursive", *Langue française*, 65, pp. 80–91.
- Zwicky, A. (1984): "Clitics and Particles", *Working papers in Linguistics*, 29, Columbus Ohio, pp. 148–173.
- Zwicky, A. M. (1985): "Clitics and particles", *Language*, 61, p. 283–305
- Žaža, S. (1979): "K úloze tzv. větných adverbii při vyjadřování postoje mluvčího k propozičnímu obsahu", *SaS*, 40, pp. 135–138.