Studii de lingvistică 7, 2017, 67 - 88

Le corpus multilingue InterCorp : nouveaux paradigmes de recherche en linguistique contrastive et en traductologie

InterCorp multilingual corpus: new research paradigms in contrastive linguistics and in translation studies

Olga Nádvorníková[1]

Abstract: Linguistic corpora have changed the research paradigm in many linguistic disciplines, allowing researchers to exploit large data based on real occurrences in authentic contexts. Using a large multilingual corpus (InterCorp, available online), this paper aims to present the changes of research paradigms brought on by corpora in two domains: contrastive linguistics and translation studies. On the basis of research results obtained by using this corpus, we demonstrate new research possibilities offered by parallel corpora, in particular bi-directional (multidirectional) analysis and the study of the specific features of the language of translation (translation universals).

Key words: parallel corpus, contrastive linguistics, translation studies, causative, constructions, gerund, punctuation, introductory verbs.

1. Introduction[2]

Les corpus linguistiques ont changé la donne dans de nombreux domaines linguistiques, permettant aux chercheurs d'exploiter de larges bases de données d'occurrences concrètes dans des contextes authentiques. La plupart des langues sont désormais dotées de larges corpus unilingues accessibles en ligne *via* des interfaces sophistiquées dont les fonctionnalités permettent aux chercheurs de trier ces données et d'y effectuer des analyses statistiques avancées (par exemple *British National Corpus* pour l'anglais, *Corpus del Español*, *Narodowy korpus języka polskiego* ou le *Corpus national tchèque*). L'apparition de

[1] Université Charles, Prague, Faculté des Lettres, Institut d'Études Romanes ; olga.nadvornikova@ff.cuni.cz.

[2] Cette étude a été réalisée grâce au soutien financier obtenu pour le projet *Progres 4, Language in the shiftings of time, space and culture* (Ministère de l'éducation nationale tchèque).

corpus parallèles (multilingues, voir ci-bas), dans les années 1990, a ensuite élargi ce champ de recherche à la comparaison entre deux (ou plusieurs) langues.

L'objectif de cet article est de présenter, moyennant un large corpus multilingue nommé InterCorp (Rosen & Vavřín 2016), disponible en ligne (cf. chap. 2), les changements de paradigmes de recherche apportés par les corpus dans deux domaines : la linguistique contrastive (cf. 3.1.) et la traductologie (cf. 3.2.). A partir des travaux concrets déjà effectués sur ce corpus, nous tâcherons de montrer les nouvelles possibilités de recherche offertes par les corpus parallèles, en particulier l'analyse bi-directionnelle (multidirectionnelle) et l'étude des spécificités de la langue de la traduction (les « universaux » de la traduction)[3].

2. Le corpus multilingue InterCorp

Le corpus multilingue InterCorp (Čermák & Rosen 2012) est un ensemble de corpus parallèles, contenant actuellement 40 langues[4], y compris des langues romanes. D'après l'acception largement répandue en linguistique de corpus contemporaine, **le corpus parallèle** (*parallel corpus, translation corpus*) est un corpus composé de textes originaux (sources) et de leurs traductions, le plus souvent alignés au niveau des phrases, comme dans l'exemple suivant[5] :

(FR) « Les étoiles sont belles, à cause d'une fleur que l'on ne voit pas... » (A. de Saint-Exupéry, *Le Petit prince*, 1999 [1943])

(CS) "Hvězdy jsou krásné, protože je na nich květina, kterou není vidět..." (trad. par Z. Stavinohová, 1989)

(DE) "Die Sterne sind schön, weil sie an eine Blume erinnern, die man nicht sieht..." (trad. par G. Leitgeb et J. Leitgeb, 1956)

(EN) "The stars are beautiful, because of a flower that cannot be seen." (trad. par K. Woods, 1971)

(ES) – Las estrellas son hermosas, por una flor que no se ve... (trad. B. del Carril, 2008)

(IT) "Le stelle sono belle per un fiore che non si vede... " (trad. par N. Bompiani Bregoli, 1997)

(RO) – Stelele sunt frumoase datorită unei flori pe care nimeni nu o vede... (trad. par B. Corlaciu, 1998)

[3] Pour la présentation détaillée des principes de la constitution du corpus InterCorp, voir Nádvorníková 2016 et Rosen & Vavřín 2016; pour l'analyse des problèmes méthodologiques liés à l'exploitation des corpus parallèles en général, voir Nádvorníková 2017a.
[4] ar, be, bg, ca, cs (langue pivot), da, de, el, en, es, et, fi, fr, he, hi, hr, hu, is, it, ja, lt, lv, mk, ms, mt, nl, no, pl, pt, rn, ro, ru, sk, sl, sq, sr, sv, tr, uk, vi.
[5] Exemples repris du corpus InterCorp (Čermák & Rosen 2012).

Cette définition du corpus parallèle est acceptée tant par les traductologues (Baker 1995 : 230 ou Laviosa 2002 : 36) que par les linguistes (McEnery, Xiao & Tono 2006 : 47 ; Altenberg & Granger 2002 : 8 ; Hunston 2002 : 15). Le corpus parallèle peut être soit unidirectionnel, c'est-à-dire contenant seulement les traductions de la langue A vers la langue B, soit bi-directionnel, à savoir couvrant les deux sens de la traduction (A ↔ B). Le corpus InterCorp est un corpus parallèle bi-directionnel (ou multidirectionnel), ce qui représente un atout méthodologique important (cf. 3.1.)[6].

Parmi les corpus multilingues, on compte également **les corpus « comparables »** (*comparable corpora*) (cf. Altenberg & Granger 2002 : 8 et Chlumská 2014). Contrairement aux corpus parallèles, ils ne se composent pas de textes originaux et de leurs traductions respectives et ne permettent donc pas l'alignement. Les corpus comparables composés uniquement de textes originaux évitent également les problèmes méthodologiques qui sont parfois liés aux corpus parallèles (les interférences de la langue source, les universaux de la traduction, etc. – cf. 3.2. et Nádvorníková 2017). Cependant, leur constitution doit soigneusement respecter des paramètres identiques concernant la taille et la composition, pour assurer leur « comparabilité » (on parle de *sampling frame*, cf. McEnery & Hardie 2012). Ce type de corpus est souvent utilisé pour l'extraction de terminologies bilingues, mais il trouve aussi d'autres applications (par exemple l'analyse des discours politiques en deux langues d'une époque donnée, cf. Lewis 2005).

Les corpus comparables de traductions (*comparable translation corpora*) servent à l'analyse de la langue de la traduction ; ils contiennent donc des textes originaux et des textes traduits (tout comme les corpus parallèles), mais dans *une même* langue (ils sont donc unilingues). Les sous-corpus de textes traduits et non-traduits (originaux) doivent être comparables quant à leur taille et leur composition. Pour l'anglais, il existe par exemple *The English Comparable Corpus* (EEC, Laviosa-Braithwaite 1996) ; pour le tchèque, l'Institut du Corpus national tchèque offre ce type de corpus à côté du corpus parallèle InterCorp (Chlumská 2013). Quant au français, à notre connaissance, un tel corpus n'est pas encore disponible, bien qu'il soit d'une grande utilité pour les recherches traductologiques, comme nous le verrons dans 3.2.

[6] Les corpus parallèles alignent en général l'original à une seule traduction dans la langue donnée (c'est aussi le cas d'InterCorp) ; parmi les exceptions à cette tendance, citons par exemple le corpus *Kačenka* (Parallel Corpus of English and Czech Texts, www.phil.muni.cz/angl/kacenka/kachna.html) ou *Kapradi*, contenant par exemple plusieurs variantes de traductions tchèques des pièces de théâtre de W. Shakespeare (www.phil.muni.cz/kapradi). Il existe également un petit corpus parallèle alignant le texte français de la nouvelle *Colomba* de P. Mérimée à ses neuf traductions en tchèque (parues entre 1875 et 1975).

2.1. Le cadre institutionnel de la création du corpus InterCorp

Le corpus parallèle InterCorp est créé par l'Institut du Corpus national tchèque de la Faculté des Lettres de l'Université Charles à Prague[7] et financé par le Ministère de l'Éducation nationale tchèque. L'Institut du Corpus national tchèque a été fondé dès 1994, en vue de la création du corpus national tchèque. En 2000, ce corpus synchronique représentatif (nommé SYN2000 et contenant 100 millions de mots) a été mis en ligne, mais il ne constitue qu'un des nombreux corpus constitués depuis par cet Institut : à part les corpus représentatifs du tchèque SYN(chronique), dont la dernière version, SYN v5 (3,836 milliards de mots) a été lancée en avril 2017, l'Institut du Corpus national tchèque a également créé ou a contribué à la création de corpus diachroniques et oraux, de corpus d'apprenants (de tchèque L1 et L2 et d'anglais L2) ou d'un corpus de discours présidentiels en tchèque[8].

De plus, InterCorp[9] n'est pas le seul corpus disponible *via* l'Institut du Corpus national tchèque qui contient des langues étrangères : pour l'anglais britannique, le français, l'italien et l'allemand, l'utilisateur y trouvera les WaCky corpora (Baroni *et al.* 2009), contenant chacun plus d'un milliard de positions[10]; les corpus comparables Aranea (disponibles en cs, de, en, es, fi, fr, hu, it, nl, pl, pt, ru, sk, zh) atteignent plus d'un milliard de mots au total (Benko 2014)[11] ; et ce tableau est complété par le corpus journalistique *Est républicain* (73 millions de mots) basé sur les textes fournis par CNRTL[12].

Corpus en langues étrangères disponibles sur http://kontext.korpus.cz	Langues impliquées	Taille du corpus
Aranea (web corpora) http://sketch.juls.savba.sk/aranea_about/index.html	cs, de, en, es, fi, fr, hu, it, nl, pl, pt, ru, sk, zh	Plus d'un milliard de mots pour chaque langue
Est républicain (CNRTL) (corpus journalistique)	français	87 984773 positions
WaCky (web corpora) http://wacky.sslmit.unibo.it/doku.php?id=start	de, fr, it, uk	En nombre de positions : 1,35 mld (de), 1,35 mld (fr), 1,6 mld (it), 1,9 mld (en)
InterCorp (corpus multilingue parallèle)	39 langues (sans le tchèque)	1 460 397 000 mots

Tableau 1 : Les corpus en langues étrangères disponibles *via* l'interface KonText de l'Institut du Corpus national tchèque

[7] www.korpus.cz.
[8] http://wiki.korpus.cz/doku.php/en:cnk:uvod.
[9] http://ucnk.korpus.cz/intercorp/?lang=en
[10] La notion de *position (de corpus)* inclut non seulement les *mots*, mais également les signes de ponctuation (donc tous les éléments issus de la *tokenization*). Par ailleurs, le corpus frWac a trouvé récemment une application intéressante : Jean-Luc Manguin (2016) a relié sa base de données pour l'orthographe lexicale *Ortholexies* (https://ortholexies.greyc.fr/) directement aux concordances de ce corpus.
[11] Kratochvílová & Jindrová (2017) montrent la possibilité d'utiliser les corpus comparables Aranea pour les recherches contrastives comparant l'espagnol et le portugais.
[12] http://www.cnrtl.fr/corpus/estrepublicain/.

Tous ces corpus, y compris InterCorp, sont disponibles gratuitement, après l'enregistrement en ligne[13] qui engage l'utilisateur à ne pas se servir des corpus à des fins commerciales et à signaler toute publication réalisée grâce à ces données (voir les références sur https://www.korpus.cz/biblio). Bien que le plus grand mérite dans la création de ces corpus revienne aux membres de l'Institut du Corpus national tchèque, parmi lesquels nous trouvons des linguistes ainsi que des informaticiens, il ne faut pas oublier que d'autres établissements et organismes y contribuent de manière considérable : la Faculté de Mathématiques et de Physique de l'Université de Prague prépare certains outils du TALN (lemmatiseurs ou étiqueteurs morphosyntaxiques, par exemple) ; la préparation des textes en langues étrangères est assurée par les différents départements de linguistique de la Faculté des Lettres de la même Université, ainsi que par des universités d'autres villes en République tchèque ; et les maisons d'édition ont fourni et continuent à fournir des versions numériques de textes pour le corpus (pour plus de détails voir Nádvorníková 2016). Sans cette concertation de ressources financières ainsi qu'humaines, la réalisation d'un projet de cette taille serait impossible.

En outre, l'Institut du Corpus national tchèque encourage les applications pratiques de tous ses corpus, par exemple en organisant des stages d'initiation au travail sur corpus pour les traducteurs ou les enseignants du primaire et du secondaire[14] ou bien en rendant disponible un dictionnaire multilingue en ligne basé sur le corpus InterCorp[15].

2.2. Composition du corpus InterCorp

La composition du corpus InterCorp a déjà été présentée de manière détaillée dans Nádvorníková 2016 et 2017a (cf. note 3) ; dans ce qui suit, nous allons donc seulement rappeler les éléments pertinents pour la recherche en linguistique contrastive et en traductologie, ainsi que les modifications et les améliorations récentes.

Dans sa dernière version, lancée en 2016 (version 9), le corpus InterCorp contient plus d'un milliard et demi de mots (pour l'ensemble des 40 langues impliquées). Le corpus est réparti en six sous-ensembles, issus de projets différents :

Noyau du corpus (litt./scientif.)	**Syndicate/ Presseurop**	**Acquis communautaire**	**EuroParl**	**Sous-titres**	**TOTAL**
328 458 000	51 177 000	450 463 000	277 945 000	538 954 000	1 647 000 000
20%	3%	27%	17%	33%	100%

Tableau 2 : Composition du corpus parallèle InterCorp (40 langues), version 9 (2016) – nombre de mots

[13] https://www.korpus.cz/signup.
[14] Cf. aussi www.korpus.cz/proskoly.
[15] *Treq*, http://treq.korpus.cz/, voir Škrabal & Vavřín (2017, à par.).

Le « noyau » (*core*) du corpus, composé en majorité de textes littéraires publiés après 1950, est le sous-ensemble le plus utilisé dans le cadre des recherches linguistiques et traductologiques, parce qu'il offre la meilleure qualité de la traduction ainsi que de l'alignement (pour plus de détails techniques, Nádvorníková 2016). Par conséquent, c'est ce sous-corpus que nous avons utilisé dans les recherches présentées dans 3.1. et 3.2. Deux tiers du corpus sont représentés par les sous-titres de films[16] et par les textes issus des institutions de l'Union européenne (*l'Acquis communautaire* – Erjavec *et al.* 2005 ; et *EuroParl* – transcriptions des débats ayant eu lieu au sein du Parlement européen entre les années 2007 et 2011[17]). Les moins représentés sont les textes journalistiques[18], tirés des serveurs multilingues SYNDICATE[19] et VoxEurop[20].

Le corpus s'élargit chaque année : la version 10, qui sera lancée en juin 2017 apportera en particulier l'augmentation de la taille du noyau du corpus et une nouvelle collection, contenant 19 traductions de la Bible (ca, cs, da, de, en, fi, fr, hr, it, lt, mk, nl, no, pl, pt, ru, sk, sv, uk).

Presque toutes les langues contenues dans le corpus sont dotées de lemmatisation ainsi que du balisage morphosyntaxique (le plus souvent à l'aide de *TreeTagger*[21]) ; cependant, comme le montre la figure suivante, les langues ne sont pas représentées de manière égale dans le corpus :

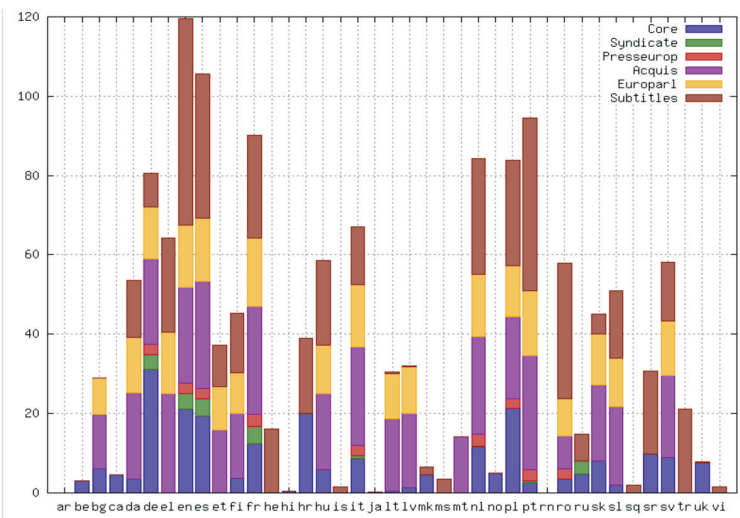

Figure 1 : Composition du corpus InterCorp en fonction des différentes langues et des sous-corpus (en millions de mots ; http://wiki.korpus.cz/doku.php/cnk:intercorp:verze9)

[16] www.opensubtitles.org.
[17] http://www.statmt.org/europarl/.
[18] Les textes journalistiques sont disponibles en : de, en, es, fr, it, nl, pl, pt, ro et cs.
[19] http://www.project-syndicate.org/.
[20] PressEurop, www.voxeurop.eu/fr.
[21] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

Parmi les langues les plus représentées dans le corpus, nous trouvons également celles qui font l'objet des études contrastives et traductologiques présentées ci-dessous : quatre langues romanes (l'espagnol, l'italien, le portugais et en particulier le français) et l'anglais.

3. Le corpus multilingue en linguistique contrastive et en traductologie

Mettant à la disposition des chercheurs de larges données authentiques, les corpus linguistiques ont rendu possibles des recherches dont la réalisation aurait été auparavant extrêmement chronophage, voire impossible. Ainsi, les analyses manuelles du gérondif (Moortgat 1978) ou des verbes introducteurs dans les incises de citation (Peprník 1969 ou Dessaintes 1960) étaient basées dans le meilleur des cas sur quelques milliers d'occurrences obtenues après un travail fastidieux, tandis que les corpus permettent de travailler immédiatement sur des dizaines de milliers d'exemples provenant de sources variées (cf. 3.1.2. pour le gérondif et 3.2.2. pour les verbes introducteurs) et d'y tester des hypothèses dans l'instantané grâce aux fonctionnalités statistiques de l'interface du corpus.

Parmi les recherches complètement inaccessibles sans les corpus, citons en particulier le sujet phare de la linguistique de corpus, les collocations (cf. par exemple Sinclair 1991). En effet, la recherche des co-occurrences et l'application immédiate des mesures d'association, permettant d'identifier la dépendance statistique de deux variables (t-score, MI score, logDice, etc.), seraient difficilement imaginables sans les corpus. De même pour la comparaison des emplois des différents signes de ponctuation et des changements dans la segmentation en phrases dans les traductions (cf. 3.2.1.). De plus, comme le remarquent Habert & Fuchs (2004 : 91), les corpus s'avèrent irremplaçables non seulement pour analyser les phénomènes fréquents (tels que la ponctuation), mais aussi dans des cas où notre intuition de locuteurs est faible, voire impuissante : les phénomènes dont la fréquence est par contre très ténue. A titre d'exemple, citons notre étude des gérondifs passés (*il se dirige vers la gauche* ***en ayant entendu*** *« promenade ! »*) effectuée sur le corpus FRANTEXT (Nádvorníková 2008) ou bien l'analyse de mots monocollocables (tels que *ado* en anglais ; ce phénomène a été étudié sur corpus en anglais, italien, allemand et tchèque dans Čermák *et al.* 2016). Sans corpus, il ne serait pas possible de rassembler un nombre suffisant d'occurrences pour en tirer des conclusions générales pertinentes.

Cependant, nous ne trouvons pas que la linguistique de corpus représente une *théorie* linguistique ; en effet, la dichotomie traditionnelle (Tognini Bonelli 2001 : 65, Tognini Bonelli 2002) entre l'approche *corpus-driven,* proche plutôt de la linguistique de corpus en

tant que théorie linguistique autonome, et l'approche *corpus-based*, concevant la linguistique de corpus en tant que méthodologie, est de nos jours mise en question. Ainsi, McEnery et Hardie (2012 : 150) proposent de ne pas considérer les approche *corpus-driven* et *corpus-based* comme opposées, mais comme des éléments d'un continuum : les deux approches sont basées sur des données empiriques et diffèrent seulement par le degré de leur utilisation (l'approche *corpus-driven* s'y fie davantage que l'approche *corpus-based*). La linguistique de corpus serait donc une méthodologie générale fournissant des données à des cadres théoriques variés (cognitifs, contrastifs, traductologiques, sociolinguistiques, etc.)[22]. Néanmoins, en se basant sur les données de corpus, ces cadres théoriques s'engagent à respecter les principes méthodologiques de la linguistique de corpus (cf. en détail Nádvorníková 2017a). C'est pour cette raison que nous ne considérons pas comme faisant partie de la linguistique de corpus les analyses utilisant les corpus seulement en tant que « bases à exemples » (*corpus-illustrated approach*) ; en effet, en séparant l'occurrence analysée de l'information sur sa fréquence dans le cadre très complexe du corpus en question, le chercheur la prive de toute fiabilité.

La question devient encore plus complexe dans le cas des corpus parallèles utilisés en recherche contrastive (3.1.), où il faut prendre en considération non seulement la fréquence de l'élément étudié (par exemple le gérondif, cf. 3.1.2.), mais également celle des différents types d'équivalents dans la langue cible. En effet, pour mettre au jour les différences et les équivalences *systémiques* entre les langues analysées, il est nécessaire de mettre de côté les équivalents particuliers et rares, et de tenter de dégager les types d'équivalents dominants (*recurrent translation patterns*, Krzeszowski 1990 : 27). Le même principe s'applique aux recherches traductologiques sur corpus : ce sont les types de changements récurrents qui nous permettent de dégager d'éventuelles spécificités de la langue de la traduction (3.2.).

3.1. Le corpus multilingue InterCorp en linguistique contrastive

L'introduction de corpus parallèles en linguistique contrastive représente un tournant méthodologique dans ce domaine : le chercheur, qui jusqu'ici fondait ses jugements sur sa compétence native en L1 et sa compétence acquise en L2, a désormais à sa disposition les données parallèles de corpus, représentant un *bilingual output* (Gast 2012). Certaines constatations contrastives traditionnelles se voient ainsi modifiées ou rejetées et d'autres différences ou similitudes sont découvertes (cf. ci-dessous) ; les comparaisons en deviennent plus

[22] Pour la méthodologie de l'application du corpus InterCorp dans le domaine de *corpus stylistics*, voir par exemple Čermáková & Fárová (2017).

objectives, nuancées et complexes. Altenberg et Granger (2002 : 7) constatent que ce sont les corpus qui ont contribué au dynamisme renouvelé des recherches contrastives ces dernières années. Parmi leurs avantages méthodologiques ils mentionnent le fait que les corpus fournissent une information plus riche et plus fiable que celle obtenue par l'introspection. Ainsi, les corpus permettent par exemple de découvrir des manières alternatives d'expression d'un sens ou d'une fonction dans la langue cible.

Par rapport aux corpus parallèles bilingues (tels que ENPC, Hansard, etc.), les corpus multilingues présentent un atout méthodologique encore peu exploré : le travail contrastif en grande équipe de chercheurs. En effet, si les corpus des différentes langues contenues dans le corpus multilingue sont suffisamment larges, ils offrent un cadre méthodologique unique au travail contrastif en équipe. Un des rares exemples de ce type d'application du corpus multilingue est le livre *Románské jazyky a čeština ve světle paralelních korpusů* (*Les Langues romanes et le tchèque à la lumière des corpus parallèles*, Čermák & Nádvorníková *et al.* 2015), qui est le fruit du travail d'une équipe de treize chercheurs en quatre langues romanes (l'espagnol, le français, l'italien et le portugais) au sein de l'Institut d'Études romanes de l'Université Charles à Prague. Le livre présente l'analyse contrastive du tchèque par rapport à quatre langues romanes dans le corpus multilingue InterCorp, en se concentrant sur quatre sujets différents : les mots complexes contenant les préfixes *re- / re- / ri- / re-* (cf. aussi Čermák 2013) et les suffixes *-ble / -ble / -bile / -vel*, les périphrases verbales ingressives, les constructions causatives *hacer / faire / fare / fazer* + *infinitif* et le gérondif/*gerundio*. Pour illustrer les changements que les corpus ont apportés à la linguistique contrastive, nous allons nous pencher plus en détail sur les deux sujets mentionnés en dernier.

3.1.1. Les constructions causatives romanes et leurs équivalents en tchèque

La causativité est une catégorie sémantico-grammaticale dont la réalisation concrète dans les langues peut être très variée. Les constructions romanes analysées (*hacer / faire / fare / fazer* + *infinitif*) relèvent de la syntaxe ; en tchèque, l'équivalent des constructions causatives traditionnellement mentionné dans les descriptions contrastives du tchèque et des langues romanes relève de la morphologie dérivationnelle – le préfixe verbal *roz-* (il s'agit donc d'un équivalent synthétique) :

(1) Ya la **hizo** llorar a su mamá – dijo la madre de la Nelly. (Julio Cortázar, *Los premios*, 1960) → "Až jste **roz**plakal svou matku", řekla Nellina matka. (trad. Blanka Stárková, 2007) (in Čermák & Nádvorníková *et al.* 2015 : 33)

Cependant, outre ce type d'équivalent, l'analyse a permis d'en identifier huit autres, dont cinq ne sont pas synthétiques, mais analytiques. L'analyse de la fréquence de ces types d'équivalents a révélé trois aspects contrastifs jusqu'ici ignorés[23] :

1) le préfixe verbal *roz-*, traditionnellement le plus souvent proposé comme équivalent des constructions causatives romanes, ne représente qu'à peu près 4% des équivalents de ces constructions en tchèque ; de plus, le nombre de verbes qui appartiennent à ce type est limité (par exemple *smát se* 'rire' et *plakat* 'pleurer') (voir (1)) ;
2) parmi les quatre types d'équivalents les plus fréquents ne se trouve qu'un seul type synthétique[24] ; les autres types sont analytiques, dont le plus fréquent n'est jamais mentionné dans les descriptions contrastives[25] ;
3) les équivalents tchèques obtenus à partir des quatre langues romanes révèlent des similitudes frappantes : les types ainsi que les proportions des quatre types d'équivalents les plus fréquents sont presque identiques (Čermák & Nádvorníková *et al.* 2015 : 81).

L'analyse du gérondif/*gerundio* a également permis de rectifier les descriptions contrastives traditionnelles, entre autres, grâce à l'application du procédé méthodologique rendu accessible par les corpus parallèles : l'analyse bi-directionnelle.

3.1.2. Le gérondif/*gerundio* roman et ses équivalents en tchèque

L'objectif de l'analyse mentionnée ci-dessus était d'identifier les facteurs qui influencent le choix du type d'équivalent tchèque de cette forme verbale romane et également de vérifier si les descriptions contrastives traditionnelles ont raison de donner comme équivalent systémique du gérondif en tchèque le transgressif (*přechodník*). Pour renforcer la rigueur méthodologique de cette analyse contrastive et sa validité pour la comparaison des systèmes linguistiques, nous avons

[23] Pour la présentation de ces résultats en d'autres langues voir également Štichauer & Čermák 2016. Pour l'analyse du même phénomène (constructions causatives) sur un autre corpus (anglais – suédois, ESPC) voir Altenberg 2002.

[24] Les verbes causatifs au sens large, liés non par un procédé de dérivation, mais uniquement par leur sens lexical (*shodit – spadnout*) : *Mais, le plus souvent, l'oisillon* ***fait tomber*** *son oeuf alors qu'il s'efforce de le briser et, du coup, s'écrase.* (B. Werber, *La Révolution des fourmis,* 1996) → *Nejčastěji však ptáče samo* ***shodí*** *své vejce, když se jím usilovně snaží proklubat, a rázem je po něm.* (trad. Š. Belisová, 2007).

[25] Il s'agit de la substitution des rôles syntaxiques: *¿Qué le hace pensar así?* → *Proč myslíte?* 'litt. Pourquoi pensez-vous cela ?' ou *Qu'est-ce qui vous fait dire cela?* → *Proč to říkáte?* 'Pourquoi dites-vous cela ?'.

choisi un *tertium comparationis* relevant de la linguistique générale : le *converb* (« a nonfinite verb form whose main function is to mark adverbial subordination », Haspelmath 1995 : 3). Pour cette même raison, nous avons limité l'analyse à l'emploi adverbial du gérondif/*gerundio*, laissant ainsi de côté le *gerundio* en construction absolue ou dans des périphrases verbales, tellement fréquentes en particulier en espagnol et en portugais[26].

L'analyse des effets de sens du gérondif/*gerundio* a révélé des similitudes intéressantes entre les quatre langues romanes : dans les quatre langues[27], l'effet de sens dominant était la circonstance concomitante (cf. Halmøy 1982 et 2003), juxtaposant simplement deux actions parallèles (Čermák & Nádvorníková *et al.* 2015)[28] :

(2) nous buvions des capuccinos […], **en regardant** la neige tomber (Jean-Philipe Toussaint, *Faire l'amour*, 2002) → my jsme pili kapucino […], **dívali se** 'nous regardions', jak před námi v uličce padá sníh. (trad. Jovanka Šotolová, 2004)

De plus, l'étude des différents effets de sens du gérondif/*gerundio* (dans son emploi adverbial) et du transgressif tchèque (cf. ci-bas) a démontré que ces formes verbales appartiennent toutes à la catégorie *contextual converb* (Haspelmath 1995 : 58, Nedjalkov 1995 : 106-110), dont le sens est vague, étant déduit seulement du contexte.

L'analyse des équivalents tchèques du gérondif/*gerundio* roman a confirmé les différences typologiques entre le tchèque et les quatre langues romanes : face à la forme verbale romane non finie, le tchèque opte pour l'expression à verbe fini (dans les traductions à partir des quatre langues, ce type d'équivalent représente plus de deux tiers des occurrences analysées, cf. (2)). De plus, la distribution des types des équivalents tchèques a confirmé la répartition des effets de sens du gérondif français proposée par Halmøy (1982, 2003) : d'une part, la circonstance concomitante, le plus souvent traduite par une proposition coordonnée (cf. (2)), d'autre part l'antériorité logique, ayant pour équivalent le plus souvent des propositions subordonnées circonstancielles, sémantiquement spécifiques (de temps, de moyen, etc.). En outre, cette correspondance des types d'équivalents aux effets de sens s'est vu confirmer pour les quatre langues romanes :

(3) **Uscendo** dalla cucina incontrammo Aymaro. (Umberto Eco, *Il nome della rosa*, 1998) → **Jak jsme vycházeli** 'Comme nous

[26] Rappelons qu'en français l'emploi adverbial du gérondif est le seul possible, si l'on ne prend pas en compte la construction progressive *aller (en) –ant*.
[27] Les échantillons analysés étaient suffisamment larges : 1448 occurrences pour le portugais, 1561 pour l'espagnol, 1862 pour l'italien et 2362 pour le français.
[28] Si le corpus n'était pas composé en majorité de textes littéraires, mais, par exemple, de textes de spécialité, l'effet de sens dominant serait plutôt 'le moyen' (cf. Nádvorníková 2012).

> sortions' z kuchyně, potkali jsme Aymarda. (trad. Zdeněk Frýbort, 1988)

Ces résultats confirment un des atouts méthodologiques des corpus parallèles observés par Johansson (2007 : 57) : les équivalents dans la langue cible nous permettent de « voir à travers les corpus parallèles » (« seeing through parallel corpora »), c'est-à-dire de distinguer les différents sens de la structure unique de la langue source (cf. également Nádvorníková 2017a)[29].

De plus, comme nous l'avons remarqué dans Nádvorníková 2017a, l'analyse des équivalents dans le corpus bi-directionnel permet d'en dégager un, considéré comme « marqueur de sens », pour chercher d'autres constructions dans la langue source appartenant au même groupe fonctionnel[30] :

> By reversing this process, *i.e.* starting from the range of variants discovered in language B and observing how these are rendered in language A, it is possible to discover paradigms of cross-linguistic correspondences (alternative ways of rendering a particular meaning or function in the target language). (Altenberg & Granger 2002: 8)

Le dernier point de l'analyse concerne le transgressif (*přechodník*) tchèque, traditionnellement donné comme équivalent systémique du gérondif/*gerundio*. Grâce à l'analyse bi-directionnelle, nous avons pu découvrir que cette équivalence n'est vraie que pour l'espagnol, l'italien et le portugais ; en français, l'équivalent le plus fréquent du transgressif tchèque n'est pas le gérondif, mais le participe présent (cf. Čermák & Nádvorníková *et al.* 2015 : 216).

> (4) Jednou nabili pana Jirouta do kanónu, a když ho vystřelili a pan Jirout dosáhl vrcholu křivky, rozpřáhl ruce a po hlavě **padaje** zvolna dolů viděl, že už dávno minul trampolínu […] (B. Hrabal, *Postřižiny*, 1976) → Un jour, on chargea M. Jirout dans son canon et lorsqu'on eut fait feu et que M. Jirout eut atteint le sommet de sa trajectoire, il écarta les bras et, **tombant** lentement, la tête en bas, il vit qu'il avait déjà dépassé le trampoline (trad. C. Ancelot, 1987)

Ce résultat démontre que l'analyse bi-directionnelle est non seulement un principe méthodologique incontournable dans l'utilisation de corpus parallèles en linguistique contrastive (cf. aussi

[29] Pour la désambiguisation des gérondifs français à travers les corpus parallèles voir aussi Nádvorníková 2013a et 2013b. Le même principe méthodologique a été appliqué par exemple dans Martinková & Janebová 2017 pour l'analyse des différents emplois de la particule tchèque *prý* (à travers ses équivalents anglais).

[30] Voir par exemple l'analyse des équivalents tchèques des verbes anglais de type *become* dans Malá 2013 et 2014.

Nádvorníková 2017a), mais également un des atouts de ce changement de paradigme que représentent les corpus en linguistique contrastive : en effet, en ajoutant la recherche allant dans l'autre sens de la traduction, on complète le tableau des équivalences systémiques des phénomènes analysés et on ajuste leur *valeur* dans les systèmes respectifs. Dans ce qui suit, nous allons montrer que l'analyse bi-directionnelle est aussi un des procédés permettant de découvrir d'éventuelles spécificités de la langue de la traduction (parfois considérée comme le « troisième code », *cf.* Baker 1998 ou Øverås 1998).

3.2. Le corpus multilingue InterCorp en traductologie

C'est déjà dans la première moitié des années 1990 que Mona Baker a remarqué le potentiel que les corpus parallèles pourraient avoir pour la traductologie (*translation studies*) ; les corpus ont changé le paradigme de recherche dans ce domaine, en permettant aux chercheurs de passer de la prescription à la description (Baker 1995 : 231 ; cf. aussi Baker 1993 et 1996). Tout comme dans le cas de la linguistique contrastive, les corpus parallèles offrent aux traductologues la possibilité de confronter leurs intuitions aux données objectives fondées sur des recherches quantitatives (grâce à la comparaison des textes traduits aux textes non-traduits, et des textes sources aux textes cibles)[31]. De plus, les traductologues ont à leur disposition des corpus comparables de traductions (cf. 2), censés également servir à la comparaison des textes traduits et non-traduits[32].

Les premières recherches traductologiques tirant profit de l'analyse des corpus s'inspiraient des travaux de Mona Baker et se concentraient sur les « universaux de la traduction », c'est-à-dire les traits spécifiques de la langue de la traduction indépendants de la langue source (Baker 1993 : 243), en particulier la simplification, la normalisation et l'explicitation. Depuis, l'objectif des recherches n'est plus de démontrer l' « universalité » de ces traits mais plutôt d'analyser leurs éventuelles manifestations dans des langues concrètes : par exemple l'explicitation en norvégien et en anglais (Øverås 1998), la réduction de la variation des cooccurents (*collocates*) dans les traductions en portugais (Dayrell 2007) ou bien la tendance à la simplification dans les traductions en espagnol (Corpas Pastor 2008) ou en tchèque (Cvrček & Chlumská 2015).

Dans ce qui suit, nous allons tenter de montrer l'utilité des corpus linguistiques en traductologie moyennant deux recherches concrètes effectuées sur le corpus multilingue InterCorp : l'examen

[31] C'est Corpas (2008) qui constate de manière explicite le changement de paradigme apporté par les corpus linguistiques en traductologie.

[32] Pour l'application du corpus Jerome sur l'analyse des spécificités de l'emploi de la ponctuation dans les textes tchèques traduits et non-traduits, voir Nádvorníková & Šotolová 2016 et Nádvorníková 2017a.

des changements de la segmentation en phrases dans les traductions impliquant le français, le tchèque et l'anglais (3.2.1.) et l'analyse des verbes introducteurs dans les incises de citations dans les traductions impliquant les mêmes langues (3.2.2.).

3.2.1. Analyse des changements de la segmentation en phrases dans les traductions impliquant le français, le tchèque et l'anglais

Les traducteurs modifient parfois la structure syntaxique du texte source, soit en reliant des phrases, soit en segmentant une phrase en deux ou plusieurs unités phrastiques (cf. (5)). Ce phénomène illustre clairement la nécessité de l'analyse bi-directionnelle (ou multidirectionnelle, cf. ci-bas). En effet, comme nous l'avons montré dans Nádvorníková 2017a (exemples (7) et (8)), l'examen de la tendance à la segmentation de phrases seulement dans un sens de la traduction pourrait indiquer des différences systémiques entre les langues analysées (sur le modèle de ce que nous avons constaté pour le gérondif et ses équivalents tchèques dans 3.1.2.). Cependant, en observant *la même* tendance à la segmentation de phrases dans l'autre sens de la traduction également, il est possible d'en déduire qu'il s'agit plutôt de l'effet d'un des universaux de la traduction (par exemple de la simplification ou de la normalisation, cf. Nádvorníková & Šotolová 2016).

Le corpus multilingue est particulièrement propice à ce type de recherche, parce qu'il permet d'explorer plusieurs sens de traduction, c'est-à-dire plusieurs paires de langues. L'analyse des traductions impliquant le français et le tchèque dans Nádvorníková & Šotolová 2016 a ainsi été complétée en y ajoutant l'anglais dans Nádvorníková 2017b. Dans cette analyse trilingue des changements de segmentation, nous avons avancé l'hypothèse que les changements de segmentation peuvent être dus soit à des différences structurelles entre les langues (le degré de la « densité informationnelle » habituelle dans les langues en question, cf. Fabricius-Hansen 1996 et 1999 ou Solfjeld 1996), soit aux tendances universelles de la langue de la traduction à la normalisation, la simplification ou l'explicitation (cf. Bisiada 2016).

L'analyse de plus de 130 000 non1:1 segments (c'est-à-dire les segments qui ne se correspondent pas quant au nombre de phrases) dans les traductions du tchèque en français et en anglais et *vice versa* a démontré que les différences concernant le degré de la densité informationnelle des langues impliquées influencent effectivement les taux de segmentation : la proportion des phrases divisées est toujours plus élevée dans les traductions en tchèque, langue à densité informationnelle moins élevée que le français et l'anglais (cf. ci-haut la tendance à l'emploi des verbes finis en tchèque, 3.1.2.). Cependant, le

taux de segmentation reste très élevé également dans les traductions à partir du tchèque (cf. Tableau 2 dans Nádvorníková 2017b), ce qui indique que les universaux de la traduction y entrent en jeu aussi, en particulier la normalisation (les phrases courtes sont reliées et les phrases longues sont segmentées) :

(5) Ale táta přece nemohl vědět, že babička umře, to jsem chápala už tehdy, a tak jsem se táty zastávala, a že nic špatnýho neudělal, si myslim i teď. (P. Hůlová, *Paměť mojí babičce*, 2002)

There was no way for Papa to know that Grandma was going to die, **though. I** realized that even **then. So** I told him he didn't do anything wrong, and I still think that today. (trad. A. Zucker, 2009)

Mais papa ne pouvait savoir que grand-mère allait **mourir. Je** le comprenais déjà à l'époque, je prenais sa défense, et je continue de penser qu'il n'a rien fait de mal. (trad. H. Rihova-Allendes et A. Maréchal, 2005) (in Nádvorníková 2017b : 12)

Dans (5), les traducteurs anglais et français ont opté pour la segmentation de la suite de propositions juxtaposées du texte source ; le texte cible devient ainsi plus clair, plus structuré que le texte source, mais l'oralité qui caractérise ce discours narratif ne se retrouve pas dans le texte cible, cette caractéristique du texte source étant effacée et normalisée. Et les corpus linguistiques nous permettent de découvrir (grâce aux données statistiques) s'il s'agit de cas isolés, ou d'une *stratégie* systématique du traducteur, modifiant le style du texte source.

3.2.2. Équivalents des verbes introducteurs d'incises de citation dans les traductions impliquant le français, l'anglais et le tchèque

Le français, l'anglais et le tchèque disposent d'inventaires de verbes introducteurs d'incises de citation très similaires, mais diffèrent quant aux normes stylistiques stipulant le degré de leur variation (*type/token ratio*) et en particulier la proportion de la répétition du verbe introducteur neutre (*say* en anglais, *dire* en français et *říci* en tchèque). Dans Nádvorníková 2017b, nous avons découvert que, dans les textes littéraires originaux du corpus parallèle InterCorp, la proportion du verbe *say* parmi les verbes introducteurs d'incises est de 61,86% ; en français, le verbe *dire* représente 51,05% des verbes introducteurs ; mais en tchèque, le verbe neutre *říci* ne représente que moins de 30% des verbes introducteurs. Ces résultats corroborent dans le cas du tchèque les recommandations stylistiques qui considèrent la répétition des mêmes verbes introducteurs comme une maladresse stylistique.

Corpus de textes littéraires originaux	Taille du corpus (en positions)	Nombre total de verbes introducteurs dans les incises	Proportion du verbe neutre (*say/dire/říci*)
Anglais	16 847 978	50 057	61,86%
Français	6 287 952	20 697	51,05%
Tchèque	18 112 612	79 573	29,17%

Tableau 3 : Proportions des verbes introducteurs neutres dans les propositions incises dans les textes littéraires originaux du corpus parallèle InterCorp

Grâce à l'analyse bi-directionnelle des textes traduits (cf. figures 1-3 dans Nádvorníková 2017c), nous avons constaté que, malgré les différences concernant le taux de répétition du verbe introducteur neutre spécifique pour ces langues, les traducteurs réussissent à respecter le taux habituel dans la langue cible. En nous basant sur ces résultats, nous avons avancé l'hypothèse que dans les traductions à partir des langues où la proportion des verbes de dire neutres (*say/dire/říci*) dans les incises est élevée vers les langues où leur proportion s'avère moins élevée, les traducteurs auront davantage recours à l'explicitation ou au remplacement du verbe de dire neutre que si la traduction se fait dans l'autre sens. Cette hypothèse s'est vu confirmer par l'analyse manuelle de l'échantillon de plus de 5 000 occurrences de verbes introducteurs dans les six sens de la traduction (cf. Figure 4 dans Nádvorníková 2017c). Par exemple, dans les traductions de l'anglais en tchèque, seulement 35,10% des *say* des textes sources sont traduits par le verbe introducteur neutre *říci*, les autres sont remplacés par des synonymes ou explicités (cf. (6)). En effet, pour arriver au taux de verbes introducteurs habituel en tchèque, le traducteur doit user de tous les moyens possibles :

(6) "That would be because they – er – weren't dementors", **said** Professor Lupin. (J. K. Rowling, *Harry Potter and the Prisoner of Azkaban,* 1999)

– C'est parce que… ce n'étaient pas des Détraqueurs, **répondit** le professeur Lupin. (trad. J.-Fr. Ménard, 2000)

To bude nejspíš tím, že – že to – nebyli mozkomorové", **zakoktal se** 'balbutia' profesor Lupin. (trad. P. Medek 2001)

L'exemple (6) montre également qu'il faut distinguer les *degrés d'explicitation* du verbe de parole dans les textes cibles. En effet, dans ce cas les deux traducteurs explicitent le verbe introducteur neutre *say* du texte source : le traducteur français précise la nature du tour de parole, évidente en contexte (réponse à une question), tandis que le traducteur tchèque, tout en rendant compte de la réalisation vocale du discours rapporté, déclenche un processus inférentiel

d'interprétation concernant les *émotions* éprouvées par le locuteur (en l'occurrence l'hésitation, la peur, l'embarras, etc.). Ce faisant, le traducteur donne au lecteur les clés d'interprétation de l'énonciation rapportée.

L'aperçu rapide de l'analyse des équivalents des verbes introducteurs d'incise dans les traductions présentée ci-dessus est très simplificateur ; cependant, il montre clairement qu'à part les différences (et les similitudes) systémiques entre les langues (cf. 3.1.) et les spécificités de la langue de la traduction (3.2.1.), les corpus parallèles (et d'autant plus les corpus multilingues) permettent d'analyser les différences concernant les normes stylistiques spécifiques pour chaque langue dont les effets peuvent être rendus visibles grâce aux larges données authentiques fournies par les corpus. Ce type de recherche représente un nouveau volet du paradigme de la recherche en traductologie.

4. Conclusion

L'objectif de cette étude était de présenter, à travers l'exemple du corpus multilingue InterCorp, les nouvelles possibilités de recherche apportées par les corpus parallèles dans les domaines de la linguistique contrastive et de la traductologie. Nous avons constaté que l'exploitation du corpus a changé le paradigme de recherche dans les deux domaines analysés, en leur fournissant de larges données authentiques disponibles à travers des interfaces sophistiquées, accessibles en ligne. Grâce à ces ressources, le chercheur peut tester dans l'instantané des hypothèses portant sur des sujets auparavant inaccessibles. Le champ de recherche s'est vu ainsi élargir au-delà de l'introspection, en particulier pour les phénomènes où notre intuition de locuteurs peut s'avérer peu fiable, à savoir les phénomènes très fréquents ou, par contre, très peu fréquents, et les données contrastives.

En effet, en linguistique contrastive, les corpus parallèles ont changé le paradigme de recherche en permettant de vérifier et, le cas échéant, de rectifier et de nuancer certaines constatations contrastives traditionnelles sur les larges données authentiques des corpus parallèles (cf. dans 3.1. les résultats de l'analyse des constructions causatives et des gérondifs en quatre langues romanes et en tchèque). De plus, l'examen des équivalents dans la langue cible jette un peu plus de lumière sur la structure analysée dans la langue source (cf. les types d'équivalents tchèques du gérondif répartis d'après ses effets sémantiques, implicites dans les langues romanes sources). Ainsi, ces équivalents nous permettent de « voir à travers les corpus parallèles » (cf. 3.1.2.). Cependant, pour obtenir des résultats fiables, il faut utiliser les corpus parallèles bi-directionnels.

Les corpus parallèles multilingues (multidirectionnels), tels que InterCorp, présentent deux atouts méthodologiques considérables par rapport aux corpus bilingues : premièrement, ils ouvrent la possibilité d'un travail en large équipe, basé sur une ressource unique et cohérente ; deuxièmement, il est possible de comparer plusieurs langues en même temps. Ainsi, dans le cas de la présente recherche, l'analyse contrastive de quatre langues romanes a révélé des similitudes (et des différences) intéressantes (cf. 3.1.) ; en traductologie, l'approche multidirectionnelle permet d'identifier les universaux de la traduction (3.2.1.) ou les différences concernant les normes stylistiques de chaque langue (3.2.2.). De plus, les corpus comparables de traductions, tels que Jerome, peuvent servir à l'identification d'éventuelles spécificités de la langue de la traduction.

Pour terminer, ajoutons que toute recherche se fondant sur des corpus parallèles et multilingues doit respecter des règles méthodologiques rigoureuses, concernant non seulement les facteurs qui assurent la représentativité des corpus (leur taille et leur composition), mais aussi – dans le cas des corpus parallèles – le sens de la traduction et les spécificités de la langue de la traduction. Cependant, les corpus ne représentent pas une théorie linguistique (*cf.* en détail Nádvorníková 2017a), mais un *outil* ; il revient donc au chercheur de choisir l'approche adéquate pour pouvoir profiter à fond des possibilités qui lui sont offertes.

Références bibliographiques

Altenberg, B. (2002), "Causative constructions in English and Swedish. A corpus-based contrastive study", in Altenberg, B., Granger, S. (2002), *Lexis in Contrast. Corpus-Based Approaches*, John Benjamins, Amsterdam, p. 97-116.

Altenberg, B., Granger, S. (2002), *Lexis in Contrast. Corpus-Based Approaches*, John Benjamins, Amsterdam.

Baker, M. (1993), "Corpus linguistics and translation studies: implications and applications", in Baker, M., Francis, G., Tognini-Bonelli, E. (éds), *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam-Philadelphia, p. 233-250.

Baker, M. (1995), "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", *Target*, 7/2, p. 223-243.

Baker, M. (1996), "Corpus-based translation studies: The challenges that lie ahead", in Somers, H. (éd.), *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager*, John Benjamins, Amsterdam, p. 175-186.

Baker, M. (1998), « Réexplorer la langue de la traduction : une approche par corpus », *Meta : journal des traducteurs / Meta: Translators' Journal*, 43/ 4, p. 1-10 (en ligne: http://www.erudit.org/revue/meta/1998/v43/n4/001951ar.pdf).

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009), "The WaCky

Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", *Language Resources and Evaluation*, 43/3, p. 209-226.
Benko, V. (2014), "Aranea: Yet Another Family of (Comparable) Web Corpora", in Sojka, P., Horák, A., Kopeček, I., Pala, K. (éds), *Text, Speech and Dialogue, TSD 2014, Lecture Notes in Computer Science*, vol. 8655. Springer, Cham, p. 257-264.
Bisiada, M. (2016), ""Lösen Sie Schachtelsätze möglichst auf": The impact of editorial guidelines on sentence splitting in German business article translations", *Applied Linguistics* 37/3, p. 354-376.
Čermák, F., Čermák, J., Obstová, Z., Vachková, M. (2016), *Language Periphery: Monocollocable words in English, Italian, German and Czech*, John Benjamins, Amsterdam-Philadelphia.
Čermák, F., Rosen, A. (2012), "The case of InterCorp, a multilingual parallel corpus", *International Journal of Corpus Linguistics*, 13/3, p. 411-427.
Čermák, P. (2013), « Las posibilidades de le studio o frecidas por los corpus paralelos: el caso del prefijo español *re-* », *AUC Philologia*, 2, *Romanistica Pragensia*, 19, p. 123-126.
Čermák, P., Nádvorníková, O. *et al.* (2015), « Románské jazyky a čeština ve světle paralelních korpusů [Les Langues romanes et le tchèque à la lumière des corpus parallèles] », Karolinum, Praha.
Čermáková, A., Fárová, L. (2017), "*His eyes narrowed – her eyes downcast:* contrastive corpus-stylistic analysis of female and male writing", *Linguistica Pragensia*, 28/2, p. 7-34.
Chlumská, L. (2013), *Korpus Jerome – comparable corpus of Czech translation and non-translation language*, ÚČNK FF UK, Praha (URL: www.korpus.cz).
Chlumská, L. (2014), « Není korpus jako korpus. Korpusy v kontrastivní lingvistice a translatologii », *Časopis pro modern filologii*, 96/2, p. 221-232.
Corpas Pastor, G. (2008), *Investigar con corpus en traduccion: los retos de un nuevo paradigm*, Peter Lang, Berlin & New York.
Cvrček, V., Chlumská, L. (2015), "Simplification in translated Czech: a new approach to type-token ratio", *Russian Linguistics*, 39/3, p. 309-325.
Dayrell, C. (2007), "A quantitative approach to compare collocational patterns in translated and non-translated texts", *International Journal of Corpus Linguistics,* 12/3, p. 375-414.
Dessaintes, M. (1960), *La Construction par insertion incidente*, D'Artrey, Paris.
Erjavec, T. *et al.* (2005), "Massive Multilingual Corpus Compilation: Acquis Communautaire and Totale", *Archives of Control Sciences*, 15/4, p. 529-540.
Fabricius-Hansen, C. (1996), "Informational Density: a Problem for Translation and Translation Theory", *Linguistics*, 34, p. 521-565.
Fabricius-Hansen, C. (1999), "Information Packaging and Translation: Aspects of Translational Sentence Splitting (German–English/Norwegian)", in Doherty, M. (éd.), *Sprachspezifische Aspekte der Informationsverteilung*, AkademieVerlag, Berlin, p. 175-214.
Gast, V. (2012), "Contrastive Linguistics: Theories and Methods", in Kabatek, J., Kortmann, B. (éds), *Linguistic theory and methodology* (WSK – Wörterbücher zur Sprach- und Kommunikationswissenschaft), Mouton de Gruyter, Berlin.
Habert, B., Fuchs, C. (2004), « Bilan et perspectives méthodologiques », *Le français moderne*, 72/1, p. 88-97.

Halmøy, O. (1982), *Le gérondif. Éléments pour une description syntaxique et sémantique*, Tapir, Trondheim.
Halmøy, O. (2003), *Le gérondif en français*, Ophrys, Paris.
Haspelmath, M. (1995), "The Converb as a Cross-Linguistically Valid Category", in Haspelmath, M., König, E. (éds), *Converbs in Cross-Linguistic Perspective. Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds*, Mouton de Gruyter, Berlin- New York, p. 1-57.
Hunston, S. (2002), *Corpora in Applied Linguistics*, Cambridge University Press, Cambridge.
Johansson, S. (2007), "Seeing through multilingual corpora", in Facchinetti, R. (éd.), *Corpus Linguistics 25 Years On*, Rodopi, Amsterdam-New York.
Kratochvílová, D., Jindrová, J. (2017), "Ingressive verbal periphrases in Spanish and Portuguese", *Linguistica Pragensia*, 27/1, p. 38-56.
Krzeszowski, T. P. (1990), *Contrasting Languages: The Scope of Contrastive Linguistics*, Mouton de Gruyter, Berlin.
Laviosa, S. (2002), *Corpus-based Translation Studies: Theory, Findings, Applications.* Rodopi, Amsterdam.
Laviosa-Braithwaite, S. (1996), *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*, PhD Thesis, Centre for Translation and Intercultural Studies UMIST, Manchester.
Lewis, D. (2005), « Corpus comparables et analyse contrastive : l'apport d'un corpus français/anglais de discours politiques à l'analyse des connecteurs adversatifs », in Williams, G. (éd.), *La linguistique de corpus*, Presses universitaires de Rennes, Rennes, p. 179-193.
Malá, M. (2013), "Translation Counterparts as Markers of Meaning. The Case of Copular Verbs in a Parallel English-Czech Corpus", *Languages in Contrast*, 13/2, p. 170-192.
Malá, M. (2014), *English copular verbs: a contrastive corpus-supported view*, Filozofická fakulta Univerzity Karlovy, Praha.
Manguin, J.-L. (2016), « Ortholexies, une base de données publique pour l'orthographe lexicale », *5e Congrès mondial de linguistique française, Jul 2016, Tours, France, 2016* (en ligne: https://hal.archives-ouvertes.fr/hal-01343991/document).
Martinková, M., Janebová, M. (2017), "What English Translation Equivalents Can Reveal About the Czech 'Modal' Particle *prý*: A Cross-Register Study", in Aijmer, K., Lewis, D. (éds), *Yearbook of Corpus Linguistics and Pragmatics. Volume 5: Contrastive Analysis of Discourse-pragmatic Aspects of Linguistic Genres*, Springer.
McEnery, T., Hardie, A. (2012), *Corpus Linguistics: Method, Theory and Practice*, CUP, Cambridge.
McEnery, T., Xiao, R., Tono, Y. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge, London.
Moortgat, B. (1978), *Participe et gérondif. Étude de l'opposition entre la présence et l'absence de EN devant la forme en –ant* (thèse de doctorat), Université de Metz, Metz.
Nádvorníková, O. (2008), « Gérondif passé - mort ou vivant? » in Albert, S. *et al.* (éds), *Le passé dans le présent, le présent dans le passé*, JATEPress, Szeged, p. 275-283.

Nádvorníková, O. (2012), *Korpusová analýza faktorů sémantické interpretace francouzského gérondivu* (thèse de doctorat), Filozofická fakulta Univerzity Karlovy, Praha.

Nádvorníková, O. (2013a), « *– Paul se rase en chantant, dit-il en bafouillant* : Quels types de manière pour le gérondif en français ? », *AUC Philologia*, 2, *Romanistica Pragensia*, 19, p. 31-44.

Nádvorníková, O. (2013b), « Les gérondifs antéposés : quelles relations avec les contextes de gauche et de droite ? » *Verbum*, 35/1-2, p. 161-174.

Nádvorníková, O. (2016), « Le corpus multilingue InterCorp et les possibilités de son exploitation », in Trotter, D., Bozzi, A., Fairon, C. (éds), *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux.* Nancy, ATILF/SLR (en ligne : http://www.atilf.fr/cilpr2013/actes/section-16.html).

Nádvorníková, O. (2017a), « Pièges méthodologiques des corpus parallèles et comment les éviter », *Corela*, HS-21 (en ligne : http://corela.revues.org/4810).

Nádvorníková, O. (2017b), "Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus », in Janebová, M., Martinková, M. (éds), *Language Use and Linguistic Structure, OLINCO 2016 Proceedings*, Palacký University Olomouc, Olomouc, p. 445-461.

Nádvorníková, O. (2017c), « Les proportions des verbes SAY/DIRE/ŘÍCI dans les propositions incises et leurs équivalents en traduction : étude sur corpus parallèle », *Linguistica Pragensia*, 27/2, p. 35-57.

Nádvorníková, O., Šotolová, J. (2016), « Změny v segmentaci na věty v překladových textech: analýza dat z francouzsko-českého paralelního korpusu [Modifications de la segmentation en phrases en traduction: analyse des données du corpus parallèle français-tchèque] », in Čermáková, A., Chlumská, L., Malá, M. (éds), *Jazykové paralely*, ÚČNK/NLN, Prague, p. 188-235.

Nedjalkov, V. P. (1995), "Some Typological Parameters of Converbs", in Haspelmath, M., König, E. (éds), *Converbs in Cross-Linguistic Perspective. Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds*, Mouton de Gruyter, Berlin-New York, p. 97-137.

Øverås, L. (1998), « In Search of the Third Code: An Investigation of Norms in Literary Translation », *Meta: Translator's Journal*, 43/4, p. 557-570.

Peprník, J. (1969), "Reporting Phrases in English Prose", *Brno Studies in English*, 8, p. 145-151.

Rosen, A., Vavřín, M. (2016), *Korpus InterCorp, version 9*, Institut du Corpus national tchèque FF UK, Praha 2016 (en ligne: http://www.korpus.cz).

Sinclair, J. (1991), *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Škrabal, M., Vavřín, M. (2017, à par.), « Databáze překladových ekvivalentů Treq », *Časopis pro moderní filologii*, 99/1.

Solfjeld, K. (1996), "Sententiality and Translation Strategies German-Norwegian", *Linguistics*, 34, p. 567-590.

Štichauer, P., Čermák, P. (2016), "Causative constructions of the *hacer* / *fare* + verb type in Spanish and Italian, and their Czech counterparts: a parallel corpus-based study", *Linguistica Pragensia*, 26/2, p. 7-20.

Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*, John Benjamins, Amsterdam-Philadelphia.
Tognini Bonelli, E. (2002), "Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach", in Altenberg, B., Granger, S. (2002), *Lexis in Contrast. Corpus-Based Approaches*, John Benjamins, Amsterdam, p. 73-96.