ANNOTATION COMPUTATIONNELLE DES RÉSEAUX ENDOPHORIQUES DANS LES TEXTES TRADUITS : CORPUS D'APPRENANTS EN TRADUCTION

Jana PEŠKOVÁ
Université de Bohême du Sud, České Budějovice

Abstract (En): The aim of this paper is to present the corpus of non-literary texts translated from Spanish into Czech; the translations being elaborated by our translation students. We propose to evaluate the functionalities of two computational tools (Analec and Sketch Engine) in the field of visualization and annotation of coreference elements in translated texts. The electronic corpus allows us to contrast the configurations of the coreference chains in the source and target texts and to analyse subsequently how the translator learners perceive and translate the relationships between the elements of these chains.

Keywords (En): text linguistics; cohesion; endophoric mechanisms; contrastive linguistics; translation

Mots-clés (Fr) : linguistique textuelle ; cohésion ; mécanismes endophoriques ; linguistique contrastive ; traduction

DOI : 10.32725/eer.2022.006

Introduction

Dans cet article, nous présenterons notre projet de création et d'exploitation d'un corpus de textes non littéraires traduits de l'espagnol vers le tchèque. Ce projet fait partie des travaux menés dans le cadre du groupe de recherche Structures textuelles et relations discursives à la Faculté des lettres de l'université de Bohême du Sud[1]. Précisons qu'il s'agit d'un corpus d'apprenants, car il est constitué de textes traduits par nos étudiants dans des séminaires de traduction spécialisée. À ce jour, aucun autre corpus de ce type n'existe en République tchèque, du moins pas pour comparer des traductions entre les langues tchèque et espagnole.

L'enjeu principal de notre projet est double. Le premier est du ressort de la linguistique textuelle contrastive. En effet, le corpus de textes traduits permet d'observer et d'analyser la (les) manière(s) dont les traducteurs perçoivent les relations existant entre des éléments coréférentiels, et, par la suite, comment ils transfèrent ces relations d'une langue à l'autre. S'ouvre ainsi un vaste champ de recherche qui peut apporter des réponses aux questions suivantes : les traducteurs conservent-ils le même paradigme de marques de cohésion utilisé dans le texte original, ou ces marqueurs varient-ils ? Si des modifications sont apportées au texte traduit, sont-elles dues aux caractéristiques structurelles de chacune des deux langues ou dépendent-elles de l'expérience du traducteur ? Si nous comparons un texte traduit par deux traducteurs, l'un débutant, l'autre expérimenté, y a-t-il des

[1] https://www.ff.jcu.cz/cz/fakulta/vyzkumna-centra/teorie-metody-a-modely-v-soucasnem-jazykovednem-vyzkumu

différences dans la sélection des marques de cohésion ? Ou encore, serait-il parfois préférable de changer le paradigme des marques de cohésion afin d'obtenir une traduction plus naturelle dans la langue cible ?

Le deuxième enjeu principal du projet de corpus de textes traduits se trouve au niveau de la didactique de la traduction. En effet, un tel corpus constituera un outil très utile pour comparer la diversité des équivalents auxquels les traducteurs-apprenants ont recouru. Ceci facilitera par la suite l'observation de l'évolution progressive des techniques appliquées par les étudiants aux différentes étapes de leur formation. Grâce à une analyse computationnelle du corpus, des zones problématiques de la structure textuelle des textes traduits pourront être identifiées, ce qui permettra de concevoir différentes activités à utiliser dans les classes de traduction.

Le projet se trouve actuellement dans sa première phase. Plusieurs textes traduits par différents étudiants ont été rassemblés et préparés pour un traitement computationnel ultérieur ; ces textes représentent le noyau du corpus. Dans cet article, nous proposons une réflexion concernant l'utilisation de certaines applications computationnelles[2] qui permettent de visualiser la configuration des réseaux endophoriques (éléments de coréférence), dans le but de comparer ces configurations dans les textes sources (TS) et les textes cibles (TC) traduits par différents traducteurs-apprenants.

Suivant cet objectif, nous définirons d'abord (Section 2.) les bases méthodologique et terminologique de notre étude, tout en soulignant les différences typologiques entre le tchèque et l'espagnol, en particulier celles qui sont pertinentes dans le domaine du marquage de la cohésion textuelle. Puis (Section 2.1.), nous décrirons brièvement les principes de constitution de notre corpus. Enfin (Section 3.), nous appuyant sur une analyse textuelle de traductions d'apprenants, nous présenterons les outils numériques qui nous aident à visualiser et à analyser les phénomènes en question.

1. Remarques terminologiques et méthodologiques

Comme nous l'avons précisé, notre attention se portera sur une propriété précise des textes, à savoir leur cohésion. L'on sait bien qu'un texte ne saurait se résumer à un ensemble de phrases placées les unes à la suite des autres sans relations logiques entre elles. En d'autres termes, il existe dans chaque texte des éléments dont la fonction est de reprendre ce qui a été énoncé précédemment par d'autres mots, de résumer ce qui précède à travers un nouveau concept ou d'annoncer ce qui va suivre. Dans la lignée de Cuenca, nous définissons ces mécanismes comme des éléments qui « manifestent une relation entre deux éléments : l'un (A) qui fournit le sens (l'antécédent) et l'autre (B) qui est compris totalement ou partiellement en relation avec lui[3] » (CUENCA 2000 : 18). Comme le propose le même auteur (*idem* 2000 : 12), nous considérons les mécanismes de coréférence comme des manifestations fondamentales de la cohésion textuelle qui « [donnent] l'ordre au récepteur de

[2] Il s'agit concrètement des logiciels Sketch Engine et Analec (pour les détails, cf. ci-dessous).

[3] « manifiestan una relación entre dos elementos: uno (A) que pone el significado (el antecedente) y otro (B) que se entiende total o parcialmente por relación a aquél ».

chercher dans le contexte un élément qui donne du sens ou permet une identification commune, créent des réseaux de sens (réseaux endophoriques, réseaux de relations coréférentielles) qui renforcent les relations entre les constituants du texte[4] ». Nous parlons ainsi de mécanismes de coréférence endophoriques, dans lesquels nous incluons l'anaphore, la cataphore et l'ellipse et que nous distinguons des mécanismes de référence exophoriques tels que la deixis temporelle, spatiale et personnelle, qui expriment toujours une relation entre le texte et des paramètres extratextuels. Les mécanismes de référence exophoriques ne faisant pas l'objet de cet article, nous n'entrerons pas dans une présentation plus détaillée du sujet.

La typologie des mécanismes de coréférence est variée : dans une perspective très générale, ces mécanismes peuvent être subdivisés en (i) mécanismes de coréférence de nature grammaticale et (ii) mécanismes de coréférence de nature lexicale (CUENCA 2000 : 40). Leur distribution dans les textes est très hétérogène et sujette à diverses altérations, en fonction d'une série de paramètres dont les plus importants sont les types et les genres textuels.

De manière générale, la typologie des mécanismes endophoriques ne varie pas radicalement entre les deux langues comparées (CZ et ES). Ainsi, les anaphores grammaticales correspondent aux morphèmes suivants :

(i) les pronoms personnels (ES : Juan y María viven en Praga. *Él* es profesor y *ella* es médica. – CZ : Jan a Marie žijí v Praze. *On* je učitel a *ona* lékařka),

(ii) les pronoms possessifs (ES : Juan y María viven en Praga. *Sus* hijos viven en Plzeň. – CZ : Jan a Marie žijí v Praze. *Jejich* děti žijí v Plzni),

(iii) les pronoms relatifs (ES : Juan y María viven en Praga, *que* es la capital de nuestro país. – CZ : Jan a Marie žijí v Praze, *která* je hlavním městem naší země),

(iv) les adverbes (ES : Juan y María viven en Praga, *donde* les gusta mucho. – CZ : Jan a Marie žijí v Praze, *kde* se jim moc líbí.)

En ce qui concerne les anaphores lexicales, citons à titre d'exemple :

ES : Juan tiene un perro. *El perro* se llama Max. (reprise fidèle) // *El animal* se llama Max (reprise hypéronymique) // *El cachorro* se llama Max (reprise hyponymique)

CZ : Jan má psa. *Ten pes* se jmenuje Max. (reprise fidèle) // *To zvíře* se jmenuje Max. (reprise hypéronymique) // *To štěně* se jmenuje Max.) (reprise hyponymique)

Si les mécanismes sont en principe les mêmes dans les deux langues, il existe pourtant des différences typologiques majeures qui se manifestent au niveau du marquage de la coréférence. La différence la plus significative concerne la détermination nominale. En espagnol, le GN anaphorique peut être introduit par un article défini, par un déterminant démonstratif ou possessif. Le tchèque étant une langue sans articles grammaticalisés, le substantif anaphorique peut figurer nu (Jan má psa. *Pes* se jmenuje Max.), ou être introduit par un morphème démonstratif ou possessif. Non seulement le système tchèque ne dispose pas d'articles, mais les nuances qui distinguent les différents morphèmes démonstratifs anaphoriques en tchèque[5] reposent sur d'autres principes qu'en espagnol.

[4] « [dan] una orden al receptor para buscar un elemento en el contexto que le dé significado o que permita una identificación común, crea redes de significado (redes endofóricas, redes de relaciones correferenciales) que refuerzan las relaciones entre los constituyentes del texto ».

[5] Cf. PEŠEK, 2014 ; ŠTÍCHA, 2013.

1.1. Le corpus d'apprenants-traducteurs (brève présentation)

La méthodologie appliquée pour la conception du corpus de textes traduits s'appuie sur le protocole proposé par SEGHIRI (2010). Ce protocole comprend quatre phases : 1. recherche et accès à la documentation, 2. téléchargement des données, 3. standardisation et 4. stockage. Dans la phase actuelle, notre corpus comprend les traductions effectuées par les étudiants inscrits aux séminaires de traduction de notre Master des années 2019-2021 (étudiants tchèques du M2 du programme de philologie hispanique ou de traduction de la langue espagnole). Le nombre de textes du corpus augmente progressivement, à mesure que nous collectons les traductions effectuées par nos étudiants. L'une des caractéristiques du corpus que nous souhaitons mettre en avant est l'homogénéité des propriétés externes des textes. Tous les textes sont des traductions qui comprennent systématiquement (i) le texte original en espagnol (TS – texte source) et (ii) sa traduction en tchèque (TC – texte cible), c'est-à-dire dans la langue maternelle du traducteur-apprenant. Les TS sont des textes authentiques partageant les paramètres suivants :

i. Il s'agit de textes issus de pages Internet à contenu informatif ou de vulgarisation (www.muyinteresante.es).

ii. Aucun des textes sélectionnés ne dépasse 2100 caractères. Cet aspect nous semble important car les textes plus longs présentent des relations très compliquées dont la numérisation serait trop longue.

2. Démonstration et analyses

L'objectif de la démonstration que nous présenterons dans cette section est de tester les fonctionnalités de deux outils d'annotation numérique, Sketch Engine[6] et Analec (version 1.4)[7], afin d'évaluer leurs performances et leur utilité pour l'analyse des relations endophoriques dans notre corpus de traductions d'apprenants. Cette évaluation est basée sur l'analyse comparative de sept traductions différentes d'un texte source (TS), effectuées par nos étudiants en master de traductologie. Le TS, relevant du style journalistique et ayant une fonction informative-descriptive prédominante, s'intitule « *Un nanochip que detecta cáncer en fase inicial* » et vient de la revue *Muy interesante* qui l'a publié en ligne le 21 mai 2014[8]. Nous reproduisons ci-dessous le texte dans son intégralité, la disposition de paragraphes est celle de l'édition originale :

> Un nanochip que detecta cáncer en fase inicial
>
> ¿Y si pudiera diagnosticarse un tumor cuando éste está en sus fases tempranas cuando apenas afecta a unas pocas células? Un equipo de científicos del Instituto de Ciencias Fotónicas (ICFO) ha logrado crear un diminuto chip que detecta marcadores de proteínas de cáncer en la sangre. Este descubrimiento evitará tener que esperar a que el tumor sea detectable a nivel macroscópico (cuando ya está formado por millones de células cancerosas), que es la fórmula general por la que se detectan la mayoría de los cánceres actuales.

[6] *Cf.* : https://www.sketchengine.eu/

[7] *Cf.* : https://www.lattice.cnrs.fr/ressources/logiciels/analec/

[8] https://www.muyinteresante.es/innovacion/articulo/un-nanochip-que-detecta-cancer-en-fase-inicial-531400667835

> El funcionamiento de este nanochip que mide apenas unos centímetros cuadrados, es sencillo: el dispositivo tiene la capacidad de detectar concentraciones muy bajas de estas proteínas de cáncer en la sangre, por lo que permite el diagnóstico de la enfermedad en una etapa precoz, lo cual puede suponer un gran avance para un tratamiento más adecuado y temprano de esta enfermedad, no solo por su fiabilidad, sensibilidad y bajo coste, sino también a su manejo y portabilidad.
>
> El dispositivo lab-on-a-chip contiene además varios sensores distribuidos en una red de microcanales de fluidos, que permite llevar a cabo múltiples análisis: monitorea cualquier cambio que se produzca en la sangre, proporcionando una evaluación directa y fidedigna del riesgo del paciente a desarrollar un determinado cáncer.
>
> "Lo más fascinante del descubrimiento es que somos capaces de detectar concentraciones extremadamente bajas de esta proteína en cuestión de minutos, lo que hace este dispositivo una herramienta de última generación, un instrumento ultra-sensible y poderoso que mejorará la detección temprana y el seguimiento del tratamiento de cáncer", afirma Romain Quidant, líder del proyecto.

À la suite d'une analyse thématique effectuée selon les principes de l'approche danešienne (DANEŠ 1994), nous avons constaté que le texte comporte trois éléments thématiques centraux dont chacun développe une chaîne coréférentielle respective : *el nanochip* [thème A – Chaîne1_Nanochip], *descubrimiento* [thème B – Chaîne2_Descubrimiento], et *el cáncer* [thème C – Chaîne3_Enfermedad]. Tout au long du texte, ces référents sont repris par une série variée de mécanismes relevant des procédés anaphoriques grammaticaux ou lexicaux (*cf.* supra section 1). Les maillons des chaînes correspondent ainsi aux syntagmes nominaux définis (anaphores lexicales), introduits par un morphème déterminatif de catégories différentes : articles définis (*el dispositivo*), articles indéfinis (*un diminuto chip*), démonstratifs (*este descubrimiento*) ou déterminants nuls (*Ø cáncer*), ou aux différents morphèmes grammaticaux (pronoms et terminaisons verbales).

Ci-dessous nous citons un fragment de ce texte (figure 1). Les différentes chaînes coréférentielles ont été annotées manuellement et « en surface du texte », sous le logiciel PowerPoint. La couleur verte correspond à la chaîne 1, la couleur rouge à la chaîne 2 et la couleur bleue à la chaîne 3.

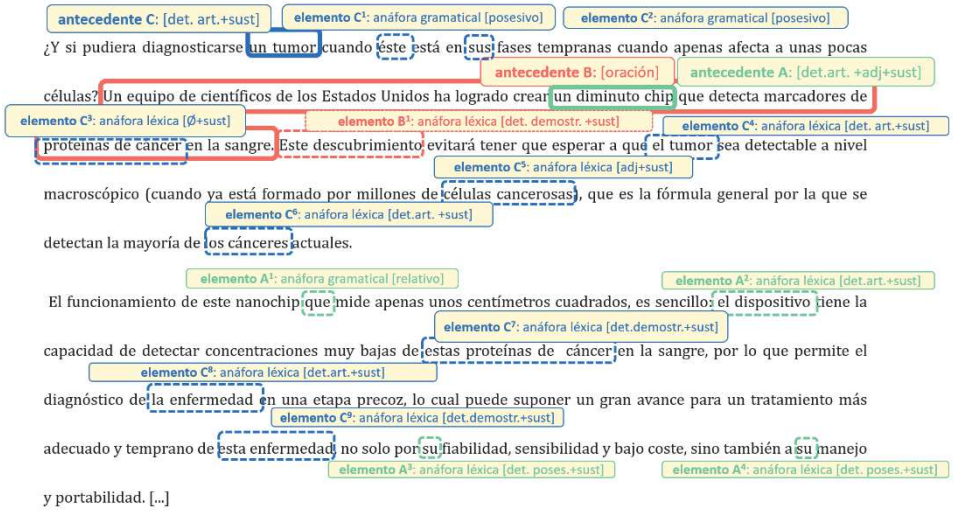


Figure 1 : Visualisation des chaînes endophoriques (marquage manuel, PowerPoint)

Bien que cette annotation permette une lecture aisée et synthétique des relations endophoriques, sa création s'avère complexe et, dans le cas d'un grand nombre de textes, demande énormément de temps. De plus, cette annotation « de surface » ne permet pas d'exploitations ultérieures, de nature statistique en particulier. C'est pourquoi il s'avère nécessaire d'avoir recours à d'autres outils numériques, qui permettent une annotation « profonde », exploitable par des applications computationnelles. Suivant cet objectif, nous avons testé deux outils d'annotation, Sketch Engine et Analec (*cf.* ci-dessus), qui offre chacun des fonctionnalités différentes.

2.1. Sketch Engine : importation, standardisation et stockage des données

Tout d'abord, les textes analysés ont été importés manuellement : soit à partir des pages Internet (TS) correspondantes, soit à partir des fichiers Word enregistrés par les étudiants sur la plateforme d'apprentissage Moodle (TC). Tous les TS ont été alignés manuellement avec leurs TC respectifs et sauvegardés au format xlsx. Ensuite, un corpus multilingue a été créé dans Sketch Engine, intitulé Traducciones_estudiantes_2021 (ES) / Překlady_studenti_2021 (CZ). La figure 2 présente la visualisation des concordances parallèles de l'un des antécédents analysés – l'antécédent C – *nanochip*. Les textes sont classés par phrases ou paragraphes (segments de texte), ce qui nous permet d'observer directement plusieurs traductions d'un segment de texte donné. En revanche, il n'y a pas de fonction d'affichage des textes dans leur ensemble. La colonne de gauche affiche un segment de TS et celle de droite les différentes traductions de ce segment dans chacun des TC.

Figure 2 : Visualisation des concordances parallèles de l'antécédent A du Texte 3 : Chaîne 1_Nanochip

2.1.1. Sketch Engine : traitement des données et visualisation des résultats obtenus

Parmi les nombreuses fonctions de Sketch Engine, aucune ne nous permet de visualiser l'ensemble du texte. Il en résulte qu'il serait difficile, voire impossible, d'y annoter les réseaux endophoriques de l'ensemble du TS et de les comparer à ceux de leurs TC respectifs. Pourtant, cet outil présente quand même quelques intérêts pour notre recherche. En effet, Sketch Engine permet de visualiser le texte analysé en fonction des segments qui le constituent. Une telle visualisation peut être utilisée pour une annotation manuelle *a posteriori* (par exemple au format *PowerPoint*) de réseaux endophoriques spécifiques. À titre d'exemple, deux référents du TS (*nanochip* et *cáncer*) et leurs équivalents respectifs dans chacun des TC analysés sont marqués dans la figure 3. Cette représentation nous permet de nous concentrer sur au moins deux aspects : d'une part (i) sur la variété des solutions appliquées par les traducteurs lors de la traduction des réseaux endophoriques identifiés dans le TS ; d'autre part (ii) sur la typologie des équivalents que les différents traducteurs choisissent pour transmettre la fonction des déterminants des syntagmes qui constituent ces réseaux endophoriques. En l'occurrence, dans l'exemple présenté à la figure 3, ce sont les articles définis (déterminants) : *el dispositivo* [DET+SUST]. Dans ce qui suit, nous allons brièvement commenter ces observations.

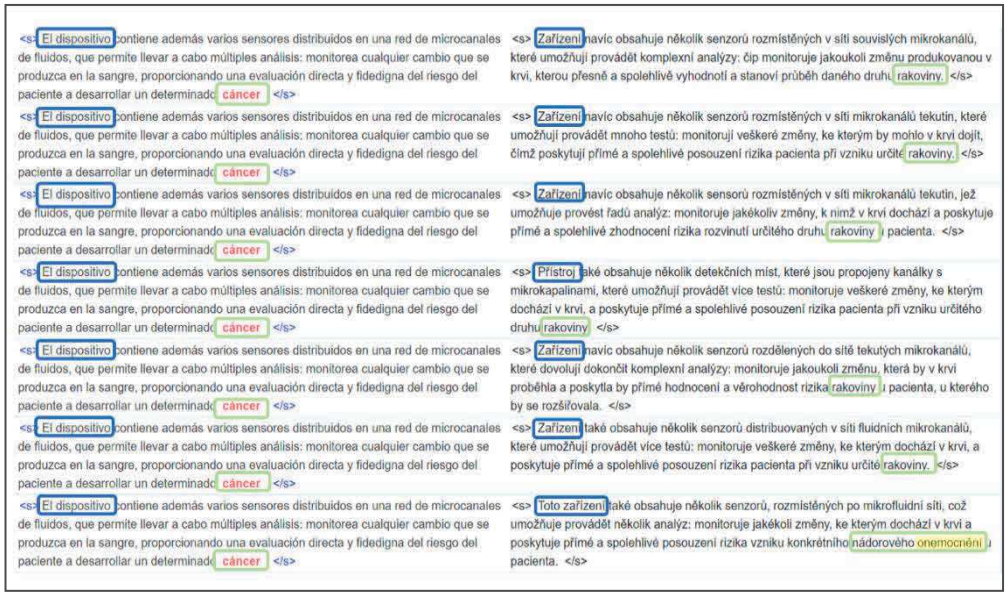


Figure 3 : Référents « nanochip » et « cáncer » dans le TS et sept TC différents annotés manuellement au format pptx.

(ad-i) En ce qui concerne la variété des solutions utilisées par les traducteurs pour reprendre ces référents (figure 4), l'analyse révèle que dans 6 cas sur 7, la même ressource endophorique que dans le TS a été retenue, en l'occurrence la

répétition totale. Cette observation semble confirmer notre hypothèse selon laquelle les traducteurs, surtout au début de leur carrière professionnelle, ne perçoivent pas le texte dans son intégralité. En d'autres termes, ils ont tendance à traduire « mécaniquement », appliquant la démarche mot à mot, sans prêter suffisamment attention à la cohésion des grandes lignes thématiques. Ainsi, ils n'exploitent pas toutes les possibilités qui existent dans la langue cible pour signaler la coréférence dans un texte et qui peuvent, en l'occurrence, différer des préférences constatées dans la langue source[9]. Cet aspect, qui mériterait évidemment une analyse plus approfondie, sera davantage commenté ci-dessous.

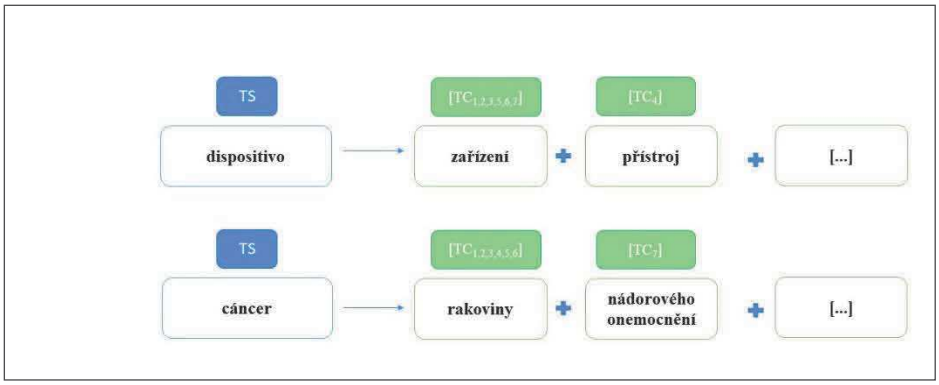


Figure 4 : Variété des solutions utilisées par les traducteurs pour reprendre les référents

(ad-ii) La figure 5 présente la variété des déterminants des syntagmes nominaux dans l'anaphore lexicale. Dans tous les TC apparaît une tendance à ne pas rendre explicite la fonction des déterminants utilisés dans les chaînes anaphoriques. À titre d'exemple, nous pouvons prendre le cas des référents de la Chaîne 1_Nanochip, et plus précisément, l'exemple du deuxième paragraphe : il apparaît ainsi que, parmi les solutions proposées par les sept étudiants/apprenants en traduction, un seul a choisi de traduire le défini par le démonstratif *toto*.

[9] *Cf.* à ce sujet LUNDQUIST (2005).

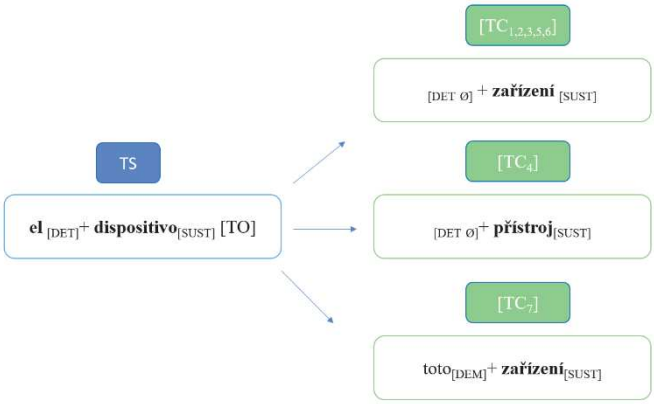


Figure 5 : Variété des déterminants des syntagmes nominaux dans les anaphores lexicales

L'analyse des solutions traductologiques montre ainsi à quel point il est important d'inclure dans la formation de futurs traducteurs les notions de linguistique textuelle : en tchèque, les SN anaphoriques nus ne fonctionnent pas de la même manière que les SN anaphoriques introduits par le démonstratif *toto*, qui véhicule des nuances sémantiques particulières (*cf*. PEŠEK 2014).

Comme nous l'avons mentionné ci-dessus, Sketch Engine permet d'identifier les équivalents des expressions anaphoriques dans le TS et dans le TC respectivement. Après cette identification, il est possible de procéder à une représentation globale, qui visualise le réseau endophorique dans son ensemble. Ces visualisations, rappelons-le, ne se font plus directement dans Sketch Engine, qui ne sert ici que de support, mais ont été ajoutées dans un logiciel de traitement d'images.

Ainsi, la figure 6 présente le schéma le plus fréquent dans notre corpus, soit celui où la typologie des mécanismes endophoriques du TS varie très sporadiquement par rapport à celle utilisée dans les TC. La densité des constituants des chaînes endophoriques entre les TS et les TC est très similaire, les procédés anaphoriques utilisés sont de même nature (en l'occurrence, il s'agit d'un segment de la Chaîne1_Nanochip).

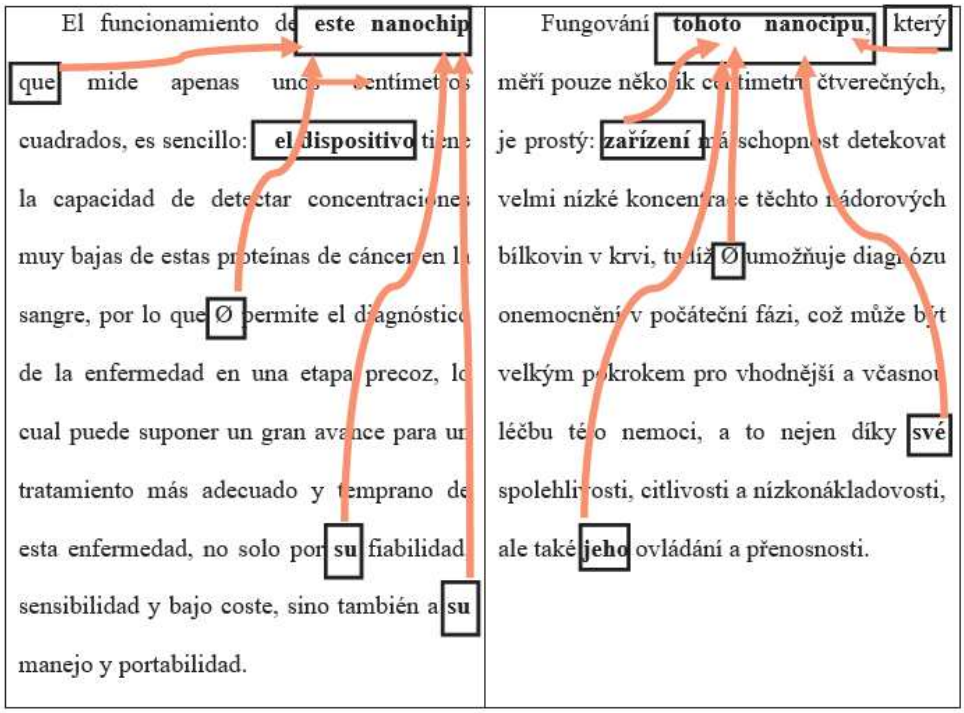


Figure 6 : Visualisation parallèle de TS et TC_1, alignés en xlsx. avec l'annotation parallèle de la chaîne endophorique « nanochip ». Exemple de chaîne endophorique structurellement identique

En revanche, dans la figure 7, nous voyons que le traducteur a appliqué une série de procédures qui aboutissent à la modification des procédés anaphoriques dans la traduction par rapport à l'original. Dans ce texte, le traducteur a opté pour une répartition différente de maillons coréférentiels par rapport à l'original : aux deux possessifs anaphoriques du TS correspondent un SN anaphorique et deux anaphores grammaticales (terminaisons verbales). Comme nous l'avons observé ci-dessus, les traducteurs débutants restent en général plus fidèles aux TS et ont une tendance plus accrue à traduire mot à mot. Dans ce sens, le TC_2 témoigne d'une démarche traductive plus avancée[10]. D'un point de vue didactique, il nous semble intéressant d'observer à quelles étapes de la formation du traducteur, ou dans quels types de traductions, ces modifications de la cohésion textuelle commencent à se manifester plus régulièrement. Il est possible de concevoir par la suite de nouvelles activités d'apprentissage qui visent précisément l'acquisition des compétences textuelles et la mise en pratique de ces compétences lors de l'activité traduisante. Nous voyons donc ici tout intérêt de la constitution du corpus d'apprenants en traduction et de son exploitation.

[10] Observons que les deux solutions, TC_1 et TC_2, sont correctes du point de vue de la cohésion textuelle. La deuxième nous semble toutefois plus élégante.

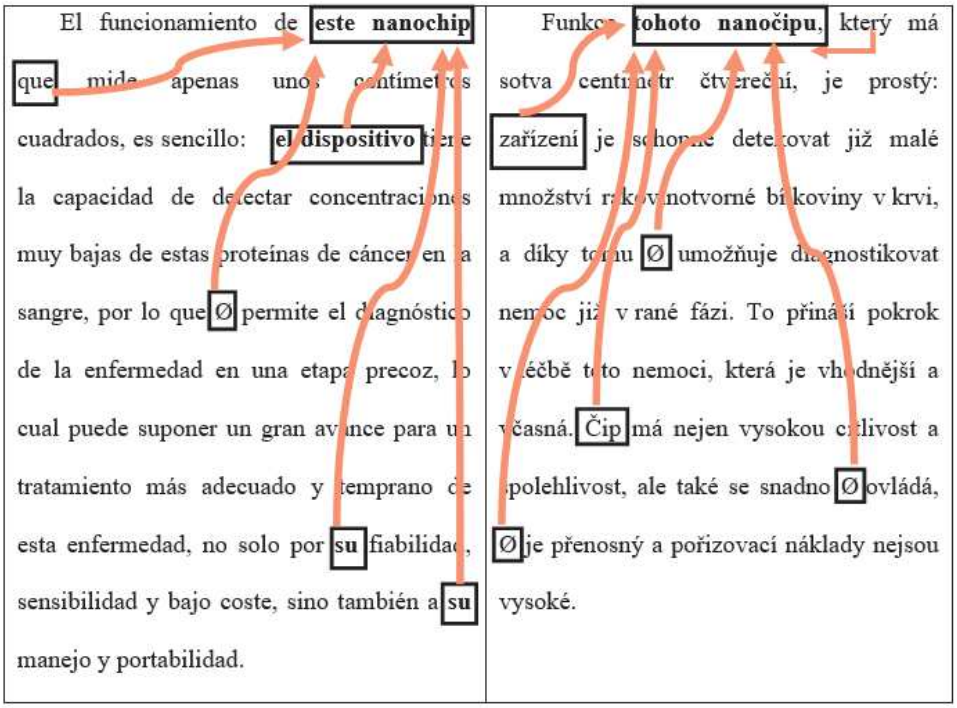


Figure 7 : Visualisation parallèle de TS et TC_2, alignés en xlsx. avec l'annotation parallèle de la chaîne endophorique « nanochip ». Exemple de chaîne endophorique structurellement différente

2.2. Analec : importation, standardisation et stockage des données

Conformément aux objectifs de cette étude (section 3), l'analyse qui suit concerne les possibilités de représentation des réseaux endophoriques à l'aide du logiciel Analec. Cet outil permet entre autres de visualiser l'ensemble du texte et de personnaliser l'annotation en fonction des objectifs du chercheur.[11]

[11] LANDRAGIN, POIBEAU, VICTORRI (2012). Ce programme a déjà été utilisé dans nos recherches précédentes, dans lesquelles nous nous sommes consacrée à l'analyse des relations anaphoriques dans les productions écrites des étudiants de ELE (espagnol langue étrangère) (2019), ainsi qu'à l'analyse contrastive des fonctions textuelles des démonstratifs (2020).

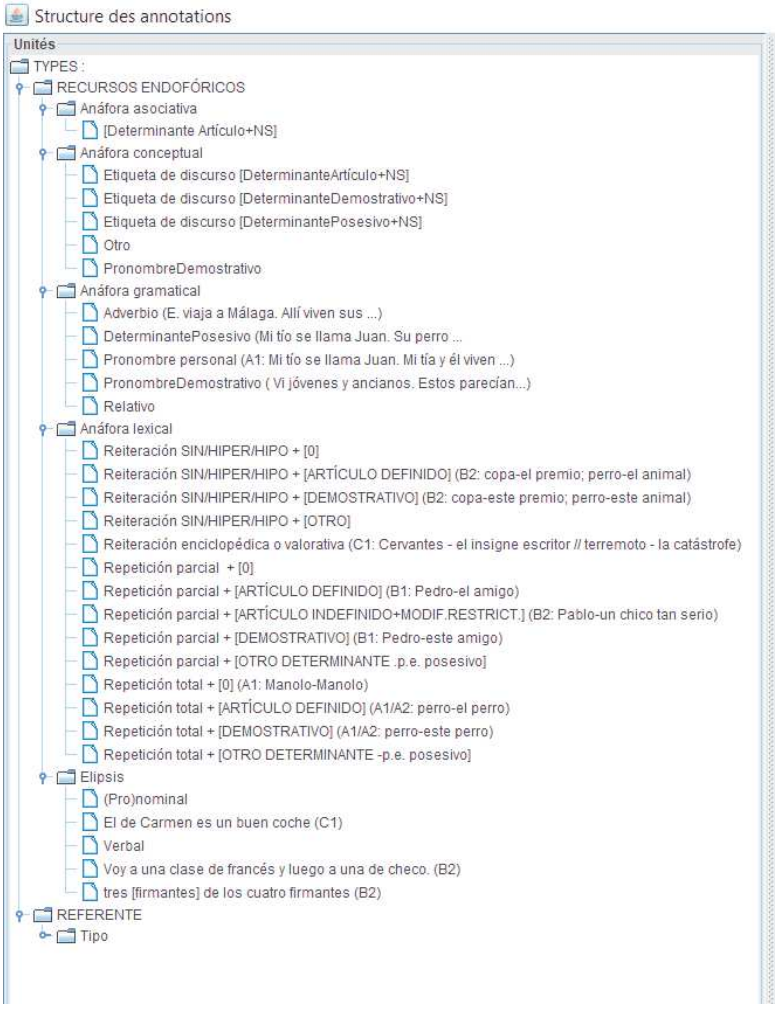


Figure 8 : Structure des annotations des maillons coréférentiels dans Analec (Unités)

Pour notre analyse, nous avons utilisé les mêmes textes que ceux présentés dans la section précédente. Dans un premier temps, les textes (texte source + textes cibles) ont été enregistrés au simple format *txt*. Dans la phase suivante, les textes ont été importés dans le logiciel Analec et les chaînes coréférentielles qu'ils contiennent (Chaîne1_Nanochip ; Chaîne2_Descubrimiento ; Chaîne3_Enfermedad) ont été annotées selon le principe Unité-Relation-Schéma[12]. La figure 8 montre l'exemple d'une structure d'annotation (niveau Unités), que nous

[12] Cf. *Manuel de TXM. Annotation URS (Unité-Relation-Schéma) version 1.0*, p. 2 : « Dans un modèle d'annotation Unité-Relation-Schéma (URS), les *Unités* (ou entités) portent sur une séquence contiguë de mots. […]. Les *Relations* quant à elle[s] relient deux éléments du modèle, et ont un nombre quelconque de propriétés (relation de type 1-à-1). Enfin, les *Schémas* contiennent des éléments du modèle, et ont un nombre quelconque de propriétés (relation de type 1-à-n). » https://docplayer.fr/174040157-Manuel-de-txm-annotation-urs-unite-relation-schema-version-1-0.html

avons appliquée aux différents maillons des chaînes analysées. L'annotation rend compte de différents procédés anaphoriques utilisés.

Après avoir annoté les différents maillons (Unités) selon la structure susmentionnée, nous les avons regroupés en chaînes (dans le cadre de l'élément Schéma), qui correspondent aux trois lignées thématiques identifiées, *cf.* figure 9 :

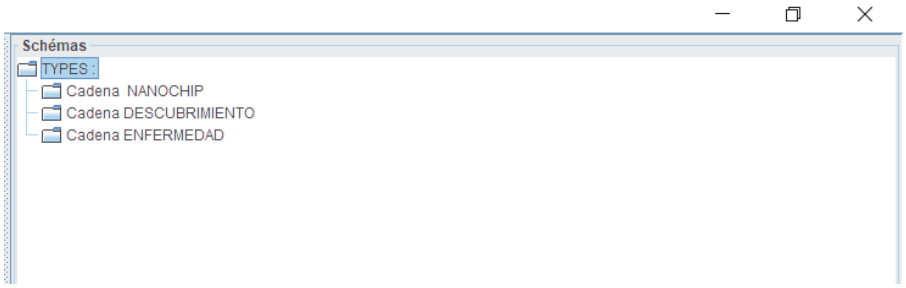


Figure 9 : Structure des annotations des chaînes coréférentielles dans Analec (Schémas)

Il est important de rappeler que chacun des textes (TS, TC_1, TC_2) a été annoté séparément, la corrélation des données entre les textes source et cibles ne se fait qu'ultérieurement, une fois le procédé d'annotation terminé.

Le logiciel Analec dispose d'une fonctionnalité qui permet de visualiser la répartition des différents maillons d'une chaîne au sein des paragraphes du texte. En comparant ces visualisations, nous pouvons constater les différences entre les textes source et cibles qui émergent ainsi d'une manière très nette. La figure 10 offre ainsi une visualisation de la Chaîne 1_Nanochip. Nous pouvons observer que le TC_1 garde la même répartition des éléments coréférentiels dans le cadre des paragraphes que le texte source (traduction plus « fidèle »), alors que le TC_2, par rapport au TS, multiplie les maillons au sein des paragraphes 3 et 4[13]. Le traducteur a ainsi restructuré le texte par rapport à l'original tout en gardant le sens communiqué ; cette solution témoigne d'un niveau plus avancé des compétences de traduction.

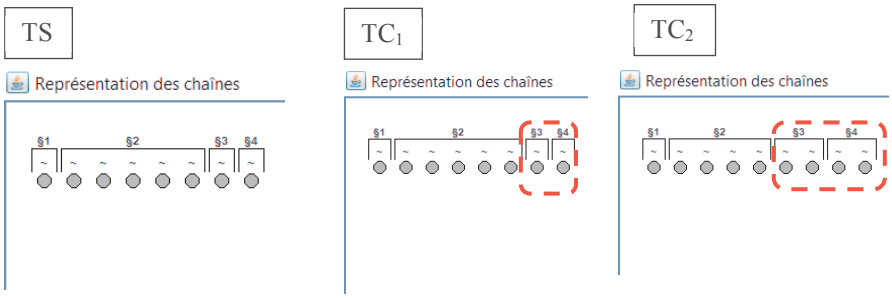
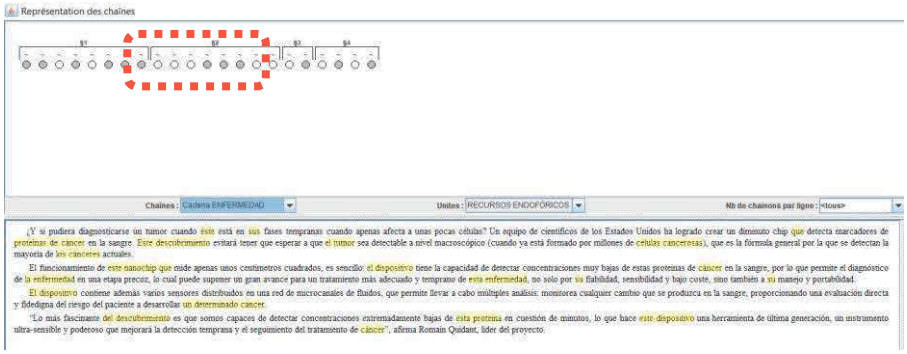


Figure 10 : Visualisation de la Chaîne 1_Nanochip dans le TS, le TC_1 et le TC_2

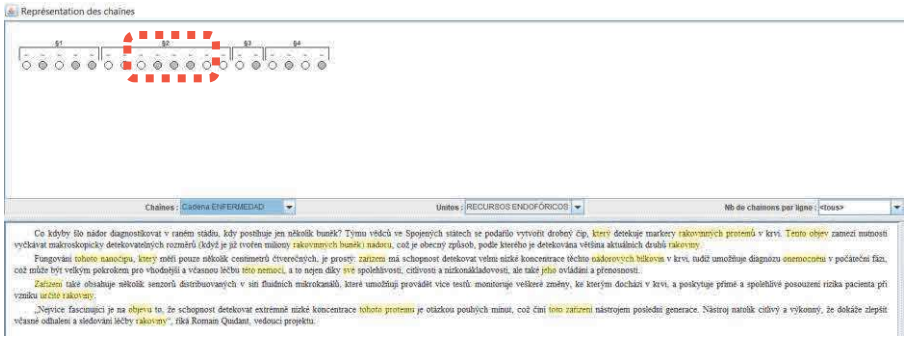
[13] S'ouvre ainsi un volet de recherche très intéressant pour la traductologie, à savoir la comparaison des corrélations thématiques entre les paragraphes du TS et du TC. *Cf.* à ce sujet ACHARD-BAYLE & PEŠEK (2020).

Si les différences constatées ci-dessus peuvent être imputées aux choix traductologiques, motivés pour la plupart par des considérations stylistiques, notre corpus annoté comporte d'autres occurrences où les non-correspondances peuvent être expliquées par les différences typologiques entre les langues comparées. Ainsi, sur la figure 11, nous observons que dans le texte espagnol, l'antécédent *tumor* (Chaîne2_Enfermedad) est repris par un pronom démonstratif *este* (*tumor* → ***éste*** → *sus [fases]* → *cáncer*). L'usage du pronom démonstratif dans cette fonction est extrêmement rare en tchèque et, si c'est le cas, cet usage signale un style administratif très marqué. Ainsi expliquons-nous les différences entre le TS et les deux TC :

TS



TC_1



TC_2

Figure 11 : Visualisation du TS et des TC_1 et TC_2 montrant la répartition des éléments co-référentiels de la Chaîne2_Enfermedad au sein des paragraphes

Enfin, l'annotation des chaînes coréférentielles nous permet de quantifier l'usage de différents procédés anaphoriques des TC par rapport au TS. Grâce aux fonctionnalités du logiciel Analec, nous pouvons ensuite évaluer les traductions du point de vue de la cohésion textuelle. Comme nous l'avons fait observer à plusieurs reprises, une traduction mot-à-mot, qui reproduit fidèlement les procédés anaphoriques du texte source, est considérée comme moins « avancée » que celle qui prend en compte les différences typologiques et les habitudes stylistiques de la langue cible. La figure 12 montre l'exemple d'une telle quantification rendue par le logiciel Analec :

TS

Résultats

	Nb d'occurrences	Fréquence (%)
Adverbio (E. viaja a Mála...	0	0.0
DeterminantePosesivo (...	2	10.53
Pronombre personal (A1:...	0	0.0
PronombreDemostrativo ...	0	0.0
Relativo	2	10.53
<aucune valeur>	15	78.95
TOTAL	19	100.0

TC_1

Résultats

	Nb d'occurrences	Fréquence (%)
Adverbio (E. viaja a Mála...	0	0.0
DeterminantePosesivo (...	2	10.53
Pronombre personal (A1:...	0	0.0
PronombreDemostrativo ...	0	0.0
Relativo	2	10.53
<aucune valeur>	15	78.95
TOTAL	19	100.0

TC_2

Résultats

	Nb d'occurrences	Fréquence (%)
Adverbio (E. viaja a Mála...	0	0.0
DeterminantePosesivo (...	0	0.0
Pronombre personal (A1:...	1	4.76
PronombreDemostrativo ...	0	0.0
Relativo	2	9.52
<aucune valeur>	18	85.71
TOTAL	21	100.0

Figure 12 : Données statistiques relatives aux procédés anaphoriques utilisés dans les TS et TC rendues par le logiciel Analec

Conclusion

Dans cet article, nous avons exploré les possibilités offertes par différents outils informatiques pour annoter et modéliser les relations de cohésion textuelle dans un corpus d'étudiants en cours d'élaboration à l'Institut d'études romanes de l'Université de Bohême du Sud. Sur la base de plusieurs textes traduits issus de ce corpus, nous avons tenté d'analyser les différences spécifiques, au niveau de la cohérence textuelle, qu'il est possible d'identifier au cours du processus de traduction. Nous avons observé l'usage de mécanismes phoriques différents dans les textes sources et cibles, qui résulte soit de divergences systémiques entre les langues, soit de l'inexpérience des traducteurs. C'est la raison pour laquelle nous avons souligné la nécessité d'inclure, dans la formation de futurs traducteurs, des éléments de linguistique textuelle et discursive. Cette dimension textuelle des traductions nous semble être quelque peu négligée et, par la suite, elle est souvent absente de l'évaluation de la qualité des textes traduits.

Nous sommes persuadée que les outils computationnels peuvent être d'une grande utilité pour l'analyse des traductions et pour l'évaluation de leur qualité au niveau de la cohésion textuelle. Notre analyse comparative a fait ressortir les avantages et les inconvénients de deux outils disponibles. Les avantages et les apports de ces outils sont indiscutables et ont été démontrés ci-dessus. Il serait certes possible d'améliorer leurs fonctionnalités en vue d'une analyse traductologique, notamment en ce qui concerne la visualisation simultanée des textes source et cible(s). Mais étant donné l'évolution rapide dans le domaine de l'analyse de la langue assistée par l'ordinateur, ces améliorations, croyons-nous, ne se feront pas attendre.

BIBLIOGRAPHIE

CUENCA Maria Josep (2000), *Comentario de textos: los mecanismos referenciales*. Madrid, Arco/Libros, S.L.

DANEŠ František (1994), Odstavec jako centrální jednotka tematicko-kompoziční výstavby textu (na materiále textů výkladových), *Slovo a slovesnost*, 55/1, p. 1-17.

SEGHIRI Miriam (2010), Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *Revista de lingüística teórica y aplicada* 49(2), p. 13-30.

CALVI, Maria Vittoria et al. (2010), Las lenguas de especialidad en español, Roma, Carocci.

LANDRAGIN Frédérix, POIBEAU Thierry, VICTORRI Bernard (2012), ANALEC: a New Tool for the Dynamic Annotation Textual Data, in *International Conference on Language Resources and Evaluation*, May 2012, Istanbul,Turkey, p. 357-362.

LUNDQUIST Lita (2005), Noms, verbes et anaphores (in) fidèles. Pourquoi les Danois sont plus fidèles que les Français, *Langue française* 2005/1 (n° 145), p. 73-91.

NEKULA Marek (2017), Koheze, in: KARLÍK Petr, NEKULA Marek, PLESKALOVÁ Jana (éd.), *CzechEncy – Nový encyklopedický slovník češtiny*, URL: https://www.czechency.org/slovnik/KOHEZE.

TORNER CASTELLS Sergi, LÓPEZ FERRERO Carmen, MARTÍN PERIS Ernesto (2011), Problemas en el uso de las anáforas en producciones escritas de español como lengua extranjera, *Revista Española de Lingüística* 41(2), p. 147-174.

PASTOR Pilar (2017), *La deixis locativa y el sistema de los demostrativos*, Madrid, Arco/ Libros, S.L.

PEŠEK Ondřej (2014), Nominální anafora a determinace – kontrastivní analýza francouzských a českých systémových možností, *Časopis pro moderní filologii* 96/2, p. 147-164.

REAL ACADEMIA ESPAÑOLA (2009), *Nueva gramática de la lengua española: Morfología. Sintaxis I.*, Madrid, Espasa Libros.

ŠTÍCHA František (2013), *Akademická gramatika spisovné češtiny*, Praha, Academia.