CLINA
Vol. 5-2, 77-98
December 2019
eISSN: 2444-1961
DOI: https://doi.org/10.14201/clina2019527798

Unique Items and Parallel Corpora: Evidence from Czech

El corpus paralelo como herramienta para explorar los elementos únicos en el checo

Michaela Martinková y Markéta Janebová
Palacký University Olomouc
michaela.martinkova@upol.cz / marketa.janebova@upol.cz

Recibido agosto-2019. Revisado: octubre-2019. Aceptado: noviembre-2019.

Abstract: This study makes a contribution to the discussion of one candidate for a translation universal, i.e. the hypothesis concerning «unique items» (Tirkkonen-Condit 2002, 2004). We address one line of criticism of this hypothesis, namely «problems with defining unique items *a priori*» (Chesterman 2007, 11). We argue that candidates for «unique items» can be revealed through Johansson's contrastive method of systematically studying «meaning through translation patterns» in a parallel corpus (Johansson 2007a), especially by comparing correspondences of a polyfunctional or vague item in source and target texts. Having previously investigated the correspondences of the Czech polyfunctional particle *prý* in English (Martinková and Janebová 2017), we now turn to Spanish. The paper touches upon problems which have to be dealt with in such contrastive parallel corpus-based studies.

Key words: Unique item; parallel corpus InterCorp; Czech particle *prý*; contrastive analysis; Czech; correspondence.

Resumen: Este estudio tiene por objeto contribuir al debate acerca de la hipótesis de elementos únicos (Tirkkonen-Condit 2002, 2004) como un candidato para un universal de traducción. Nos enfocamos en los problemas con la definición de elementos únicos *a priori* (Chesterman 2007, 11), que constituyen un punto de crítica al respecto. Partimos de la tesis de que los elementos únicos se pueden revelar usando el método contrastivo: Johansson (2007a) muestra que el significado y la función de un elemento se pueden estudiar de manera sistemática a través de su traducción en un corpus paralelo, comparando las correspondencias de un elemento polifuncional o vago en los textos origen y los textos meta. Tomando el caso de la partícula polifuncional checa *prý*, estudiamos las correspondencias españolas en textos paralelos en InterCorp (corpus paralelo del checo y otros idiomas), reflexionando sobre los problemas que surgen en tales estudios contrastivos basados en corpus paralelos, como por ejemplo la calidad del checo como una lengua de menor difusión.

Palabras clave: Elemento único; corpus paralelo InterCorp; partícula checa *prý*; análisis contrastivo; checo; correspondencia.

1. INTRODUCTION

Since 1993, when Baker predicted «the elucidation of the nature of translated text as a mediated communicative event» to be «the most important task that awaits the application of corpus techniques in translation studies» (Baker 1993, 243), researchers in the field of translation have been struggling to find out whether «universal features of translation» (*ibid*.) really exist. Arguably, the study of what Chesterman calls T-universals is even more pressing for small languages, where translations represent a high proportion of published material[1], reaching a wide readership of native speakers of these languages and possibly exerting influence on them (Chlumská and Richterová 2014a, 17). Translated texts have, for example, traditionally been included in the SYN corpora (corpora of contemporary written Czech) of the Czech National Corpus.

We want to contribute to the discussion of one candidate for a universal, i.e. the hypothesis concerning «unique items» (Tirkkonen-Condit 2002, 2004). First, we will briefly describe the main reasoning behind the hypothesis and summarise the critical notes that the hypothesis received. We will then address one line of criticism, namely «problems with defining unique items *a priori*» (Chesterman 2007, 11). While Chesterman (*ibid*.) calls for «contrastive corpus studies to investigate which items manifest significantly different frequencies in translations […] vs. non-translations», we will argue that candidates for «unique items» can also be revealed through Johansson's contrastive method of systematically studying ambiguity and vagueness «through translation patterns» (2007a), especially as a result of comparing correspondences in

[1] According to Chlumská and Richterová (2014a, 17), 34 % of the books published in the Czech Republic in 2012 were translations.

source and target texts. The paper will also touch upon methodological issues which have to be addressed in a contrastive parallel corpus based study of the present type.

1.1. *Unique items hypothesis and its critical appraisal*

In her attempt to identify «the linguistic features shared by texts assumed to be translations, as well as those shared by texts assumed to be originally produced», Tirkkonen-Condit (2002, 209) noticed «that the feature that seemed to guide the subjects' decisions was the frequency vs. scarcity of target language specific (unique) items in the texts: their frequency led subjects to assume, correctly or incorrectly, that a text was original rather than translated». The author concluded that «the unique items in non-translations vs. translations deserve further research in respect of their frequency and the impressions they make on readers» (*ibid.*), and formulated a hypothesis that linguistic elements «unique in the sense that they lack straightforward linguistic counterparts in other languages» will have «lower frequencies in translated texts than in originally produced texts» (2004, 177-178). The reason for this is that «[s]ince they are not similarly manifested in the source language, it is to be expected that they do not readily suggest themselves as translation equivalents, as there is no obvious linguistic stimulus for them in the source text» (*ibid.*). Tirkkonen-Condit (2004, 178) confirms the hypothesis e.g. for Finnish verbs of sufficiency and two Finnish clitic particles, *-kin* and *-hAn*; genre effects were also observed in the sense that the difference between original and translated Finnish was more marked in Fiction than in Academic texts of her corpus.

The criticism that followed concerned two facts: the first was that the phenomenon might possibly not be unique to translation, and the second that unique items were not properly defined. As for the former, Chesterman (2007, 10) suggested that «the unique items hypothesis [...] also applies to other communication contexts in which extra difficulties are present, such as the need to speak or write in a second language». This, we believe, is a valuable observation rather than a counter-argument, recently under intense investigation: features of translated language could in fact be similar to those found in the interlanguage of L2 speakers (for the term, see Selinker 1972). Kruger (2018, 9), for example, calls for «expanding the third code», pointing out that it has long been suggested that the features of translated language

> are not unique to translations only, but are evident in a larger set of varieties characterised by diverse communicative constraints. These constraints include, amongst others, discourse production under conditions of bi- or multilingual language activation, and the relaying or «mediation» of an existing message.

The critical note concerning vagueness in the definition of a unique item is, however, serious. First, if the unique item hypothesis is claimed to be a universal, an item should perhaps be unique with respect to all other languages. Still, as Chesterman reminds us (2007, 5), «no claim is being made about the uniqueness of, say, Finnish sufficiency verbs with respect to all other human languages»; he admits that «[t]esting such a claim would indeed be quite a task». In reference to Tirkkonen-Condit (2002, 2004)[2] it is suggested that «[w]e should conclude that "unique" means "present in the target language, but not present in a similar way in a given source language"» (Chesterman 2007, 5), or in several source languages, which is also what we argue for in this paper.

Second, as Chesterman noticed, «[i]f we identify a unique item in terms of the non-existence of a straightforward, one-to-one equivalent in some other language(s), this depends in turn on what we mean by equivalence» (*ibid.*). He goes on to say that «[i]f a verb (such as a Finnish sufficiency verb, like *ehtiä*) is translated into English as a phrase (e.g. verb+object+adverb: "have time enough"), we have an instance of what Catford (1965, 79) called a unit shift» (Chesterman 2007, 7)[3]. However, there are many shifts of this kind which «do not seem to be among those suggesting the existence of unique items, in the sense described by Tirkkonen-Condit» (*ibid.*). After all, this is what we see in Špínová's 2018 study, which confirms underrepresentation of prefixed verbs in Czech translations from English (as opposed to non-translated Czech) only for one group of verbs under investigation. On the other hand, if we extend the hypothesis along the lines of Cappelle (2012, 5) («an item is more unique for a given language as its grammatical environments are more unique for that language»), we face the opposite problem: «we can say that manner-of-motion verbs are relatively unique items in English compared to French, despite the fact that when taken on their own, most of them have perfect translation equivalents» (Cappelle 202, 5).

Indeed, motion verbs are a case in point here. As Cappelle (2012, 16) found out, «manner-of-motion verbs are underrepresented in English texts translated from French, a verb-framed language, but not in English texts translated from German, like English a satellite-framed language». The reason is that in Verb-framed languages (such as Romance languages) the verbal root of a motion verb encodes the path, while in Satellite-framed languages the path is encoded outside the verbal root, leaving the root to encode e.g. manner of motion. Not only are lexicons of motion verbs in these Verb-framed languages less rich than in Satellite-framed languages, but these

[2] «In the 2002 article, the context is defined as "other languages, or at least […] the source languages of the translations". In 2004, it is "other languages". In an email to me, Tirkkonen-Condit specifies that she is really focusing on the source languages of specific translations» (Chesterman 2007, 5).

[3] In contrastive corpus linguistics, correspondences belonging to a different category than the item under investigations are called «divergent», as opposed to «congruent» (Johansson 2007b, 25).

languages are also constrained in the expression of boundary crossing events: Spanish literal translations of sentences such as *The bottle floated into the cave* would not be grammatical, and manner of motion would in most cases be left unexpressed. This is reflected in the translations from a Verb-framed language to a Satellite-framed language: Martinková (2018) reveals a significantly lower range of Czech motion verbs prefixed by *v(e)-* [in] in the subcorpus of Czech translations from Spanish than in a comparable subcorpus of translations from English.

The most serious drawback of the hypothesis is what Chesterman calls «the cart before the horse», i.e. «problems with defining unique items *a priori*». That is to say, «if we are looking for instances of this category, we of course need to know in advance what we are looking for» (Chesterman 2007, 11). As a way out of this loop, Chesterman (2007, 11) suggests using «contrastive corpus studies to investigate which items manifest significantly different frequencies in translations (TT) vs. non-translations (NT)» and checking only then if the items identified as underrepresented in TTs lack direct counterparts in the other language. Such a procedure would, however, not only exclude cases such as those suggested by Cappelle, but a problem with underdeterminacy would remain: Chesterman does not explain what he means by an «item». What is needed is a purely quantitative corpus-driven research where items are defined in other than linguistic terms. For revealing chunks overrepresented in translations from certain languages, an n-gram approach, to name at least one example, has shown some interesting results (see Chlumská and Richterová 2014b).

This paper, however, takes a different route. We believe that this is actually where parallel corpus based contrastive linguistics has a lot to contribute: a systematic study of correspondences in parallel corpora can provide empirical evidence for (a type of) equivalence. That is to say, according to Gast (2012, 8, 3), high quality translations represent valuable «balanced» bilingual output, which «not only provides the empirical basis for contrastive studies but also […] can be used to establish comparability between categories from different languages». In the words of Altenberg and Granger, «[t]ranslation corpora have the advantage of keeping meaning and function constant across the compared languages» (2002, 9), i.e., original and translated texts should theoretically express what Tirkkonen-Condit[4] calls «the "same" semantic or pragmatic meaning»; arguably, this concerns linguistic units of various kinds.

Though contrastive linguists have studied «translation paradigms» for long, it is only

[4] Chesterman reports that Sonja Tirkkonen-Condit (personal communication to Chesterman) in fact suggests that one should «[s]tart from contrastive analyses of given language pairs. Select items which turn out to have the "same" semantic or pragmatic meaning, but which are formally different in the two languages (or where formal equivalents actually have different functions). Then compare the frequencies of these items in translations and non-translations» (2007, 12). According to Chesterman, however, «this method [again] needs a careful a priori interpretation of levels of formal equivalence» (*ibid*).

> [t]he use of multilingual corpora, with a variety of texts and a range of translators represented, [that] increases the validity and reliability of the comparison. It can be regarded as the systematic exploitation of the bilingual intuition of translators, as it is reflected in the pairing of source and target language expressions in the corpus texts. (Johansson 2007a, 52).

Johansson, who borrowed the term «translation paradigm» from Levenston's 1965 paper on contrastive syntax, used it not only for «forms and their possible translations» (Johansson 2007a, 52), but also for «the forms in the target text which are found to correspond to particular words or constructions in the source text» (2007a, 56). A systematic study of translation paradigms[5], arguably, can help reveal a meaning/ function of the linguistic unit in question. Malá (2012, 172), for example, uses Czech «as an auxiliary language, or a repository of translation equivalents which may serve as markers of the meaning of English copular verbs». Similarly, a detailed analysis of «translation solutions» for Spanish and Italian analytical causative constructions in Czech allows Štichauer and Čermák (2016,18) to reveal verbs «which convey[s] causativity as a lexically specified feature of the meaning» as a dominant correspondence in translation.

If the corpus is bidirectional, correspondences can be searched in both directions, i.e. sources as well as translations. In the words of Johansson, «we can ask both "How is *well* translated?" and "Where does *well* in English translated texts come from?"» (2007a, 57). This is indeed very common, especially for small languages, where more is translated into than translated from. Differences between languages in this respect are necessarily reflected in the structure of parallel corpora including a small language; subcorpora of fiction in Czech, created on the basis of the parallel corpus InterCorp (version 11, Rosen et al. 2018), for example, include 3,470,961 tokens in the Czech-to-English direction of translation, but 18,995,151 in the opposite direction.

Crucially for this study then, a detailed comparison of correspondences of polyfunctional or vague items in the source (ST) and target texts (TT) can reveal an asymmetry. In other words, one type of correspondence can have different frequencies in STs than in TTs, or be entirely missing. For example, while *I wish* followed by a finite clause was translated by the stance expression škoda že [pity that], this expression was not found to correspond to *(I) wish* in English TTs, in other words the chunk škoda že did not trigger the translator to use the phrase *(I) wish* in the dataset studied (Martinková 2014). Similarly, while the English construction with the verb *have* NP Ving with a pronoun is found to be translated with a Czech construction with a dative of

[5] Sometimes also referred to as «translation equivalents (Čermák et al. 2010), «recurrent translation patterns (Krzeszowski 1990), or «translation solutions» (Štichauer and Čermák 2016).

interest[6], this type of *have* construction is not found as a translation counterpart of Czech sentences containing these types of non-thematic datives. Such asymmetries in translation paradigms may in turn lead to different frequencies of the item in TTs than is STs: the construction *have* NP Ving with a pronoun is about three times more frequent in STs (2.65 pmw) than in TTs (0.865pmw)[7].

In the following section, we present a follow-up study of one candidate for a unique item, the Czech particle *prý*. As a starting point we take our 2017 analysis of the particle through its English correspondences.

2. *PRÝ* AS A UNIQUE ITEM

The particle *prý* can be classified as a unique item on the grounds of having no congruent (direct) translation equivalent; its dictionary equivalents come from different categories: e.g. *allegedly*, *they say*, *it is rumoured.* Furthermore, *prý* is polyfunctional: according to the Czech dictionaries, the expression has two meanings. In the first it is a modal particle with the meaning of uncertainty and doubt caused by the fact that the information is only second-hand (example [1]), in the other meaning *prý* introduces somebody else's direct speech (example [2]).

(1) Je *prý* nemocen.
[be:3SG PART ill]
"PRÝ [I hear] he is ill".

(2) Přišel k nám Jan. *Prý* dělej, jdeme do kina.
[came:PTC.M to us Jan:M.NOM. PART hurry.up:IMP, go:PRS.1:PL to cinema]
"John came to us. PRÝ [he said] hurry up, we're going to the cinema".

In our 2017 study, we examined the functions of the particle using the method of studying meaning through translation patterns. We wanted to see whether *prý* expresses uncertainty and doubt (modal overtones) in all contexts in which it does not introduce direct speech, and whether there is any difference between the genres represented in the parallel corpus InterCorp.

[6] For example, «You still have her dusting?» «Takže ona ti tu pořád uklízí?» [so she.NOM you.DAT here still cleans.PRS.3SG.F] "So she is still cleaning for you?" (InterCorp, Lindsay)

[7] However, underrepresentation of a certain type of correspondence in the ST does not have to necessarily mean that the item will be underrepresented in TTs, let alone be unique. For example, Grebeň (2019) observes that while the Czech similative demonstrative *takový* [such] is proportionally less frequent as a TT equivalent of *the sort/kind of N* than as its ST equivalent, its relative frequency in translations from English and non-translated Czech is comparable.

The first question had deeper theoretical implications: we attempted to describe the pragmatic mechanism under which the modal overtones arise. Though it is not our intention to delve into the theoretical status of the particle here, it needs to be pointed out that *prý* was originally a reporting verb (*pravit* [say]) and that Hirschová and Schneiderová (2012) were the first to call *prý* a lexical marker of evidentiality, classifying information according to its source (grammaticalised in some languages). With *prý* it is verbal report; more specifically, evidentiality which *prý* marks is – if we borrow Aikhenvald's (2004) terminology – both reported (i.e. the authorship is not specified) and quotative (i.e. the author is introduced)[8].

As for the latter aim of our 2017 study, our analysis showed not only different functions of the particle across the genres (e.g. it turned out that in fiction the dominant function of *prý* was quotative), but also revealed the fact that the particle is significantly underrepresented in Czech translations of fiction written in English as compared to original Czech fiction. This is in agreement with Špínová (2018), who includes *prý* in her analysis of Czech unique items[9]. Interestingly, however, the same could not be said about the subcorpus of Subtitles, where the difference in frequency did not reach significance; we also noted a higher percentage of correspondences with an English reporting clause introducing the original speaker in Czech TTs than in STs. We attributed this to the role of genre: in subtitles, where space is very limited, *prý* is useful to translate a whole reporting clause. Differences between correspondences of ST and TT *prý* were also revealed in fiction, where a higher percentage of evidential adverbs (e.g. *apparently*, *allegedly*, *supposedly*) was observed in translations of the particle than in English STs. Finally, and importantly for this paper, the analysis confirmed that the correspondences are divergent, not congruent, i.e. the corresponding items belong to a different category (none of them is a particle).

While the analysis suggests the importance of genres in considering potential uniqueness of items, it does not address the issue of the relative nature of uniqueness, concerning both the relativity of the notion of equivalence and of the number of languages compared. Furthermore, our study did not investigate the differences

[8] For more information about the mechanism triggering the modal overtones, as well as reflections on the difference between evidentiality and epistemic modality, see our 2017 study.

[9] Špínová (2018) compares the frequency of several candidates for uniqueness in non-translated Czech and in translations from English, using subcorpora of fiction created on the basis of the monolingual comparable corpus Jerome and the corpus of synchronic Czech SYN2015. Most of the items tested represent lexical units created by prefixation or suffixation, but also include items such as *prý*, *totiž* [because; or rather], *čili* [in other words; that is]*, nikoli* [no; but not] and forms of address which include names of professions (in Czech these are typically preceded by *pan/paní* [Mr/Mrs]), i.e., items with no direct translation counterparts in English. Arguably, the selection was done on the basis of intuition and contrastive analysis (118), which, however, does not seem corpus-based.

between correspondences in English as ST and TT language in more detail. Since language specificity and the direction of translation is also a potential drawback of the method of studying meaning through translation (Do different languages reveal the same functions of an item? Do correspondences in STs reveal the same function as correspondences in TTs?), we now compare non-translated Czech with translations from Spanish to see 1. whether there is again a difference between the frequency of the particle *prý* in Czech STs and TTs; 2. whether the same genre difference can be observed; and 3. whether Spanish correspondences differ depending on whether Czech is ST or TT. On the basis of InterCorp, version 11, we created subcorpora of Czech translations from Spanish and of Czech original texts; due to limitations given by scarcity of data and the fact that Europarl data are not systematically annotated for source language, we did this for fiction and subtitles only. Since the subcorpora of subtitles are very small, we also included the non-standard Common Czech form of the particle (*prej*)[10].

2.1. Prý/prej *in Czech STs and TTs*

The size of each of the subcorpora (two of fiction translated from and into Spanish, and two of subtitles, all aligned with Spanish), as well as the frequency (absolute and relative) of *prý/prej* in it, is presented in Table 1. Though only novels written by Spanish authors were used, the subcorpus of translations from Spanish is almost twice as large as the subcorpus of original Czech fiction. In contrast, though South-American films were included, the subcorpora of subtitles are very small.

	FICTION		**SUBTITLES**	
	Cz_ST	**Cz_TT**	**Cz_ST**	**Cz_TT**
Size in tokens	1,969,750	3,458,913	148,394	287,393
prý/prej/cumulative (AF)	546/80/626	610/2/612	11/16/27	24/3/27
prý/prej/cumulative (pmw)	277.2/40.6/317.8	176.3/0.6/176.9	74.1/107.8/181.9	83.5/10.4/93.9

Table 1. Information about Subcorpora of Czech STs and TTs – Fiction and Subtitles.

As far as the standard form of the particle *prý* is concerned, the analysis shows a picture very similar to our 2017 analysis: a significantly higher frequency of *prý* in Czech

[10] *Prej* was not included in the original study since it is stylistically restricted – it was neither expected nor found in the Europarl subcorpus and in the PressEurope subcorpus. However, it is reasonable to expect its occurrence in texts of fiction and in subtitles.

STs than in translations from Spanish (p < .001, χ2= 59.46794) in the subcorpora of fiction and not a significant difference (χ2= 0.02225) between the frequencies of *prý* in STs and TTs in the subcorpora of subtitles. When it comes to the non-standard form of the particle (*prej*), it is infrequent in the TTs of both fiction and subtitles; however, in the subtitles as STs it exceeds the frequency of *prý*, making the difference between the cumulative frequency of both forms of the particle significantly higher in STs (at p < .05, χ2= 5.42708) than in TTs even in subtitles. Still, error plots (Figure 1) created via the Lancaster Stats Toolbox do not show any significant difference between the cumulative frequencies of *prý/prej*, neither in fiction, nor in subtitles.

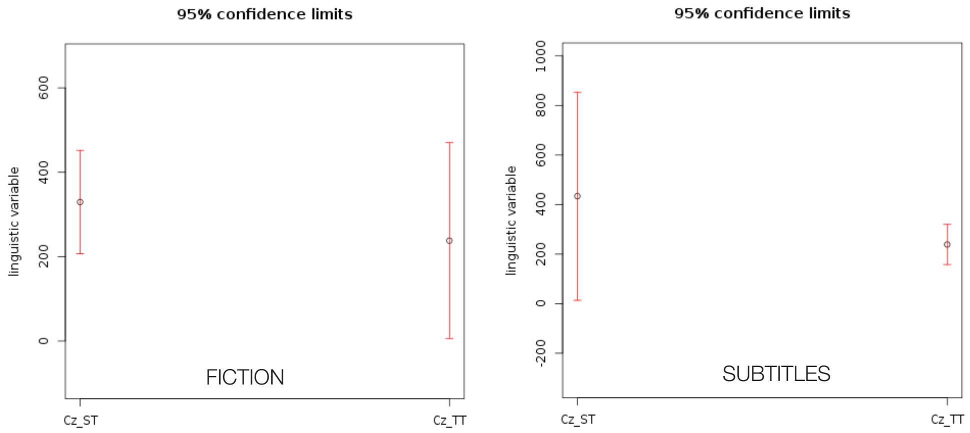

Figure 1. Prý/Prej in the Subcorpora of Fiction and Subtitles: Error Plots.

The reason is that, as Boxplot visualisations suggest (Figure 2), *prý/prej* is not evenly distributed across the texts in three of the subcorpora. In the case of fiction, there are two outliers in each of the Czech STs and TTs, and there are also two outlier texts in Czech STs of subtitles.

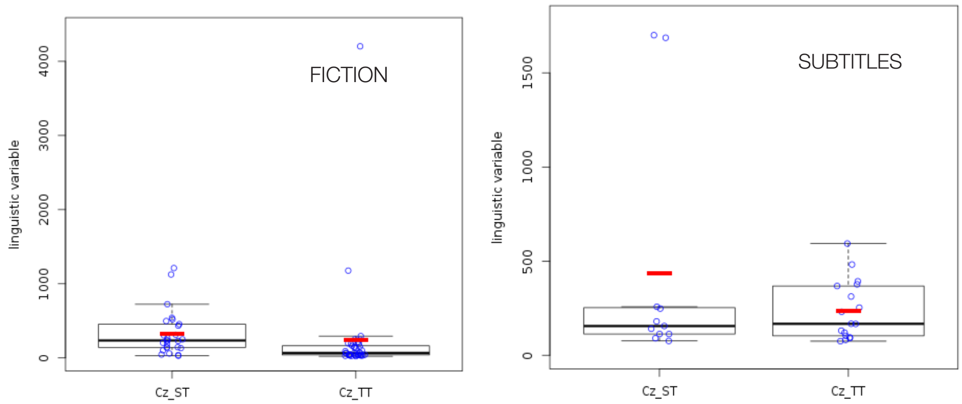

Figure 2. Prý/Prej in the Subcorpora of Fiction and Subtitles Including Outlier Texts: Boxplots.

The two outlier texts of Czech ST fiction are Milan Kundera's novel Žert (*Joke*) and Karel Čapek's *Kniha apokryfů* (*Apocryphal Tales*), where the relative frequencies of *prý/prej* are 1,125.1 pmw and 1,211.3 pmw, respectively. In TT fiction, the outlier texts are two novels by Miguel Delibes, namely *Cinco horas con Mario* (*Five hours with Mario*) (1,177.1 pmw), and *Diario de un cazador* (*Diary of a hunter*) (4,202.2 pmw). Interestingly, the last book has the highest relative frequency of the particle in all the texts included. To avoid effects of the translator's or author's style, we excluded all the outlier fiction texts from subsequent qualitative analyses of the Spanish correspondences of *prý/prej*. The following Figure 3 shows Boxplot distributions of *prý/prej* in the new subcorpora of ST and TT fiction texts as well as an Error plot showing that the difference between STs and TTs is indeed present

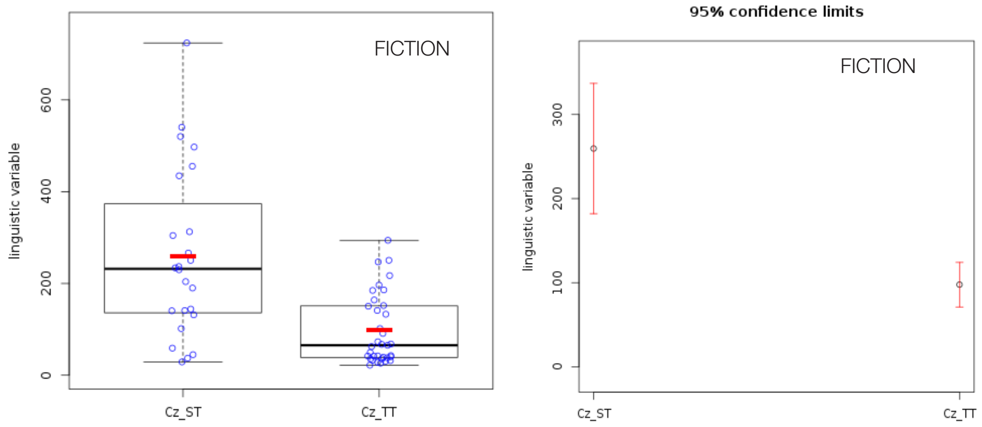

Figure 3: Prý/Prej in Subcorpora of Fiction, Outlier Texts Excluded: Boxplot and Error Plot.

However, the situation in subtitles is more complicated. As noted above, outlier texts were only found in ST subtitles: *Fimfárum* (*Fimfarum*) and *Tobruk* (*Tobruck*). Since the subcorpora of aligned subtitles are already very small (ST 148,394; TT 287,393) and outlier texts were identified only in one of the subcorpora, we could not exclude them from subsequent qualitative analysis of correspondences. For the purposes of purely quantitative comparison of the frequency of *prý/prej* in Czech ST and TT subtitles we thus created another subcorpus of non-aligned Czech ST subtitles, one which we did control for outliers. When this was done, the significance of the difference between ST and TTs was lost even for cumulative frequencies of *prý/prej*. This is confirmed by the Error plot in Figure 4 with cumulative frequencies of *prý/prej*: though no outliers were identified in TTs, the differences between individual texts are rather large. We can thus conclude that overrepresentation of the cumulative frequencies of *prý/prej* has not been proved for the subcorpora of subtitles.

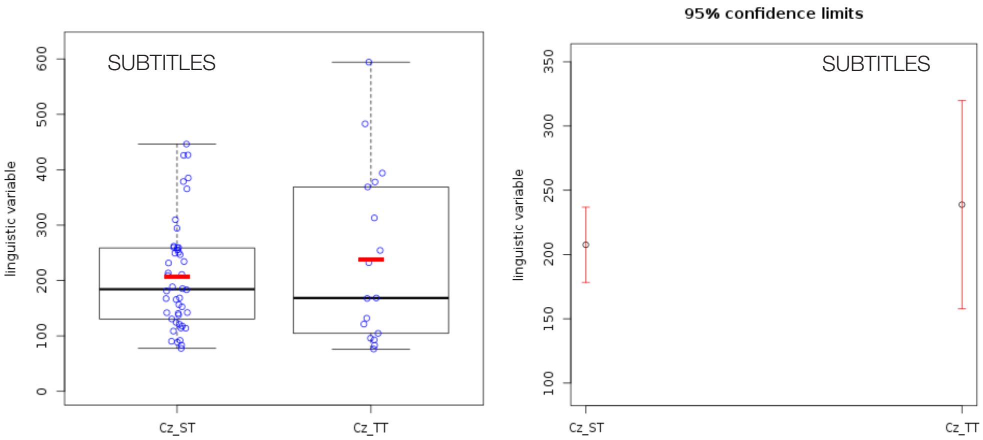

Figure 4. Prý/Prej in Subcorpora of Non-Aligned Subtitles, Outlier Texts Excluded: Boxplot and Error Plot.

Table 2 repeats information about the subcorpora of subtitles aligned with Spanish, shows the size of the newly created subcorpus of non-aligned subtitles, and provides information about the final subcorpora of fiction (excluding outlier texts) and the frequencies of *prý/prej* in them. All aligned texts were then subjected to qualitative analyses.

	FICTION		SUBTITLES		
	Cz_ST	Cz_TT	Cz_ST non-aligned	Cz_ST	Cz_TT
Size in tokens	1,819,085	3,313,889	542,620	148,394	287,393
prý/prej/ cumulative (AF)	371/78/449	273/2/275	41/33/74	11/16/27	24/3/27
prý/prej/ cumulative (pmw)	203.9/42.9/ 246.8	82.4/0.6/83	75.6/60.8/136.4	74.1/107.8/ 181.9	83.5/10.4/93.9

Table 2. Information about the Modified Subcorpora of Czech STs and TTs – Fiction (Outlier Texts Excluded) and Subtitles, and about Corpus of Non-Aligned Cz_ST of Subtitles.

In coding the data we broadly followed the procedure adopted in our 2017 study. First, it was ascertained whether it is possible to identify the original speaker (source_YES: reporting clauses with referring nouns/pronouns in the subject, expressions such as *según NP* [according to]), or not (source_NO: reporting clauses with generic subjects (*dicen* [they say]), reflexive verbs (*se dice* [it is said]), nouns equivalent to *rumour*). In Cz_STs a wider context had to be consulted, while in Czech TTs the Spanish counterparts were taken as a cue. The results are summarized in Figure 5:

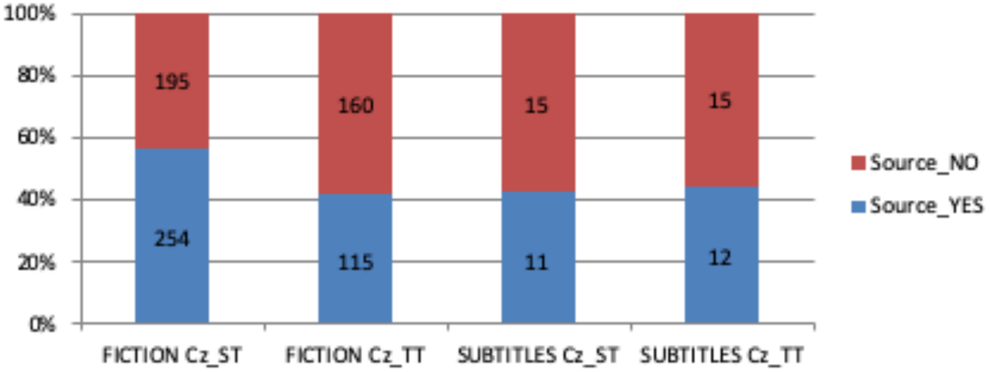

Figure 5. The Distribution of Presence/Absence of a Specific Source of Information across Subcorpora.

As Figure 5 shows, the only subcorpus in which the information introduced by *prý/prej* can be traced to a specific source in more than 50% of the cases is the subcorpus of ST fiction (56.6%); the percentage is very similar to what we observed in Czech STs aligned with English (57.4%). However, we do not see the difference observed between the subcorpora of fiction and subcorpora of subtitles: first, though in both subcorpora of subtitles the specific source of information is unknown in the majority of tokens (like in our 2017 study), the absolute number of tokens is so low that no strong conclusions can be made. Second, the subcorpus of fiction TTs shows the lowest percentage of reference to a specific source of information of all the subcorpora, which makes it not only different from our 2017 subcorpus of translations from English (where in 60% of cases there was a specific source of information), but also from the subcorpus of Czech STs. We will readdress this issue briefly at the end of the paper.

2.1.1. Spanish correspondences of the particle *prý/prej* in fiction and in subtitles

The correspondences of *prý/prej* were divided into two groups: overt and zero. While in overt correspondences a direct counterpart of *prý/prej* can be traced (in [3][11] and [4] it is a reporting clause, in [5] a prepositional phrase), in zero correspondences *prý/prej* is either added in the Czech translation or omitted in the translation into Spanish ([6] and [7] respectively)[12]:

(3) Me dijo que todo fue muy bien.
"He told me that everything went very well".

Prý bylo všechno v pořádku. (SUBT_TT:Volverás)
[PART be.PST.3SG.N everything in order]
"PRÝ everything was OK".

(4) … dicen que dijo.
"… they say that he said".

… řekl *prý*.
(FICT_TT:Caligrafía _de_los_sueños)
[say.PST.3SG.M PART]
"… he said PRÝ".

(5) Según Cristina, mi madre no la soportaba

Prý ji máma nesnáší. (FICT_TT: Amor, curiosidad, prozac y dudas)

[11] STs are presented in the left column, and translations in the right column.

[12] In our 2017 study we further sorted the tokens where *prý/prej* had no overt correspondence into indirect correspondences and zero correspondences proper. The reason was that in indirect correspondences an immediate context provided information that the information is second-hand. This was useful for English; Spanish, however, has in many cases of what was originally coded as indirect correspondence an overt correspondence between *prý/prej* and the Spanish conjunction *que*.

"According to Cristina, my mother couldn't stand her"

PART she.ACC mum hate.PRS.3SG.F
"PRÝ her mum hates her".

(6) Fue durante la noche; [me lo ha dicho uno de los guardias, que es primo de mi cuñada.]

"It was during the night; [one of the guards told me, who is my sister-in-law's cousin.]"

Prý se to stalo v noci. (FICT_TT: El maestro de esgrima)

[PART REFL it happen.PST.3SG in night]
"PRÝ it happened at night".

(7) Najdu *prý* to docela snadno. (FICT_ST:Saturnin)
[find.PRS.1SG PART it quite easily]
"I'll PRÝ find it quite easily".

[Aseguró que no podía equivocarme, que la calle era la tercera desde abajo y la casa la séptima de la izquierda.] La encontraría fácilmente.
["He said I couldn't go wrong, that it was the third street from the bottom, and the seventh house on the left.] I would find it easily"

If we consider the percentage of individual correspondences of the particle *prý/prej* in fiction, the most frequent are those with a reporting clause. The subject of the clause is either specific (quotative function, as in [3]), generic (as in [4]), or the verb is a reflexive one; the latter two cases were coded as «reported»[13]. Both functions are rather frequent, more frequent in Cz TTs than in STs: *prý/prej* translates a Spanish reporting clause with a referring noun/pronoun in the subject in 20% of tokens of TT *prý/prej* (89.9% of these contain the verb *decir* [say], other verbs are *alegar* [allege], *añadir* [add] and *contar* [tell]). ST *prý/prej* is translated by a Spanish reporting clause with a referring noun/pronoun in the subject in 17.4% (72% of these contain the verb *decir*, the rest are *alegar*, añadir, *afirmar* [affirm], *preguntar* [ask])[14].

Reporting clauses with generic subjects are even more common as correspondences of Czech TT *prý/prej* (i.e., in ST Spanish) than reporting clauses with referring nouns/pronouns in the subject; they cover 40.7% of all correspondences of TT *prý/prej*, and they translate Czech ST *prý/prej* in 12.2% of tokens of the particle*.* The reporting function of *prý/prej* in the fiction subcorpora is further confirmed by the correspondence

[13] Note again that the distinction goes back to Aikhenvald (2004). It is sometimes argued that if the source of the information is not clear, the information is potentially not reliable (hence the modal overtones).

[14] In our 2017 study the differences were larger, especially due to a high frequency of reporting clauses with specific subjects in the correspondence of ST *prý* (36.1%), and the percentage was lower for TT *prý* (21.4% tokens of TT *prý* corresponded to an English reporting clause with a specific subject).

with the conjunction *que* [that]; *que*, we would like to argue, stands here metonymically for the whole reporting clause. It is found to correspond both to ST and TT *prý* ([8] and [9], respectively):

(8) Kdo *prej* jste a jak se jmenujete. (FICT_ST:El_libro_de_los_amores_Ridiculos)

[who PART be.PRS.2PL and how REFL call.PRS.2PL]

"PRÝ who are you and what is your name".

[No paraba de hacer preguntas –se dirigió a Klara–: Más que nada preguntaba por usted.] Que quién era y que cómo se llamaba.

"[He kept asking questions – he addressed Klara – : More than anything, he asked for you.] QUE who was he and what was his name".

(9) Que qué le parecía la idea.

"QUE what did you think of the idea".

Prý jak se mu ten nápad líbí. (FICT_TT:Rabos de lagartija)

[PART how REFL he.DAT that idea like.PRS.3SG

"PRÝ how does he like the idea".

In subtitles the quotative function (correspondence with a reporting clause with a specific subject) is observed for six out of 26 tokens (23.1%) in STs and for ten out of 27 tokens in TTs (37%), i.e. we confirm the results observed in our 2017 study: in subtitles, where space is precious, a short particle is a useful means to translate a whole reporting clause. What we do not confirm, however, is a high percentage of communication verbs profiling the recipient of the message: there are just two tokens, one of *oír* [hear] and one of *escuchar* [listen to], both in correspondence of ST *prý/prej*:

(10) *Prej* to byl strašně hodnej chlapec. (SUB_ST: Tomorrow I'll Wake Up and Scald Myself with Tea)

[PART it be.PST.3SG.M terribly good boy.

"PRÝ he was a very good boy".

He oído que era un buen chico.
"I heard he was a good boy".

To answer our third question, namely whether Spanish correspondences differ depending on the direction of translation, we need to take a more systematic look at the correspondence type, and especially relative frequency of each type. First, Figure 6 brings the frequencies of overt and zero correspondences of *prý/prej* in each of the subcorpora. It suggests a difference in the percentage of zero correspondences of *prý/prej* between ST and TTs: in ST fiction it is 27%, in TT fiction only 16%; in ST subtitles

42.3%, in TT subtitles 25.9%. In other words, in both genres, the particle *prý/prej* is more often omitted in translation than added.

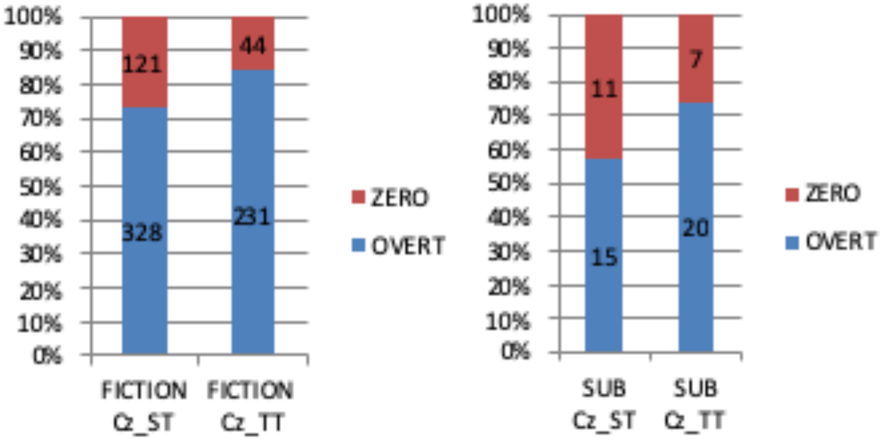

Figure 6. Distribution of Zero and Overt Correspondences of Prý/Prej in the Subcorpora.

While in subtitles the difference between the correspondence types of ST and TT *prý/prej* can be explained by a tendency not to add for reasons of space (there is also a higher percentage of correspondences with reporting clauses, both with specific and generic subjects, in TTs than STs), in fiction the situation is more difficult to explain. One of the reasons, we believe, could be the fact that *prý/prej* are frequent in reported clauses; in our 2017 paper we argued that «even if there is an evidential marker such as the verb *say*, in Czech there is a tendency to reinforce it lexically with *prý*» (2017, 89). The data seem to suggest that this type of correspondence is underrepresented in the Spanish-Czech direction of translation; note that in (11) *que* has *že* as its counterpart.

(11)	Že *prej* tě přikope až sem do pivovaru… (Fict_ST:Cutting It Short)	Que le dará un puntapié que volará hasta la cervecería [–dijo el ayudante…]
	[that PART you.DAT make.move. by.kicking.PRS.3SG as.far here to brewery]	"That he will kick you that you will fly to the brewery [– said the assistant…]"
	"That PRÝ he_will_get you_to_the_ brewery_by_kicking you"	

As to overt Spanish correspondences, striking differences between Czech ST and TT *prý/prej* can be found. This is demonstrated in Figure 7, which compares relative frequencies of Spanish correspondences of Czech ST and TT *prý/prej*; there are 130 tokens of Spanish lexical units roughly corresponding to English evidential adverbs (*supuestamente* [supposedly] and *al parecer*, *por lo visto* [apparently]) in the translations of ST *prý/prej*, while just one token of *al parecer* triggered the use of *prý* in Czech TTs.

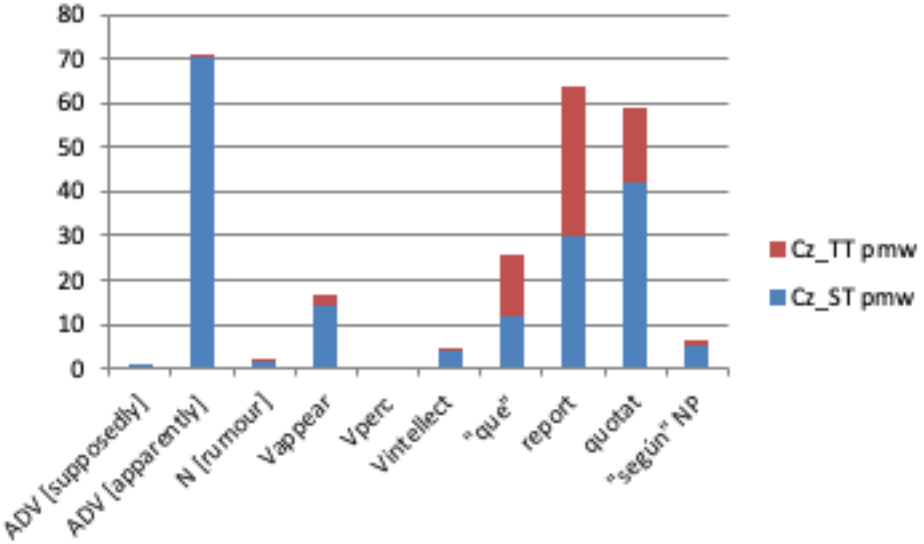

Figure 7. Relative Frequencies of Individual Spanish Correspondences Type of Czech ST and TT Prý/Prej.

Before jumping to the conclusion that we have just found a missing Spanish triggerer for *prý/prej*, (at least) one alternative explanation for a striking overrepresentation of *por lo visto* and *al parecer* as translations of Czech ST *prý/prej* has to be considered, namely the possibility that what we have here is not a case of what Gast (2012) calls a «high quality translation». This is confirmed for at least some tokens of *por lo visto*: in four novels by Jakub Urban (*Hastrman* [*El mago del agua*], *Lord Mord*, *Stín katedrály* [*La sombra de la catedral*] and *Sedmikostelí: gotický román z Prahy* [*Siete Iglesias*]), all translated by Kepa Uharte, the phrase *por lo visto* was used to translate *prý/prej* 67 times, covering 15% of all Spanish correspondences of ST *prý/prej*, and 87% of all tokens of *por lo visto* in the translation of ST *prý/prej* in the subcorpus of fiction. *Por lo visto* is, for example, used to translate 29 tokens of *prý/prej* out of 39 in *Hastrman*, and it is a dominant equivalent also in the other novels translated by Uharte. Though one might object that this is due to an author's style rather than translator's, this is not supported by evidence either: *por lo visto* is found even in sentences where a source of the reported information can be traced, i.e. a reporting clause could have been used (the function is quotative).

On the other hand, the phrase *al parecer* is found in the translations of 13 Czech books and by no means is it a dominant translation counterpart of *prý/prej* found in these books. In other words, it is a legitimate translation solution, underrepresented in correspondence of TT *prý/prej*. Furthermore, a much higher percentage of Spanish correspondences for ST than for TT *prý/prej* is also observed for verbs of appearance *parecer* and *verse*, and verbs expressing intellectual states such as *entender*, *suponer*, i.e. expressions which suggest that the truth of the presented information is open to discussion. These, we believe, could have triggered *prý/prej* in TTs more often than they did in the Spanish-to-Czech direction of translation.

3. CONCLUSIONS

Our results show that *prý* is a unique item not only for Czech as compared to English but also as compared to Spanish. It is an empirical question whether it is unique for Czech as compared to other languages, or possibly language sets, and an even broader question whether its underrepresentation is specific to translation.

Our study confirmed a significantly lower frequency of the particle *prý/prej* in Czech TTs than in STs of fiction. Unlike other studies, we pointed out as potentially problematic the fact that an item does not have to be evenly distributed across the texts, which might influence the results. The way to accommodate this problem was to exclude outlier texts. The situation with subtitles was more complicated, not only for the scarcity of data in both directions, but also because outlier texts were only found in one of the subcorpora, namely in the subcorpus of Czech STs, which, in addition, was very small. To control for outliers, we created a corpus of non-aligned Czech ST subtitles, which was larger than STs of subtitles aligned with Spanish and allowed for excluding outlier texts. No statistically significant difference between the frequencies of *prý/prej* was found, which confirms our argument that in the study of unique items genre differences have to be taken into consideration. As to the comparison of Spanish correspondences of Czech ST or TT *prý/prej*, we found a difference in the number of Zero correspondences: more specifically, in both genres *prý/prej* is more often omitted in translation than added. While this makes sense for subtitles (additions are not welcome for reasons of space), in ST fiction the particle is often used inside of reported clauses to reinforce other signals of reporting, which translators from Spanish hesitate to do.

As to overt correspondences, an absence of *por lo visto*, *al parecer*, and *supuestamente*, i.e. expressions suggesting that the information should be taken with a grain of salt, was observed in correspondence of TT *prý/prej*. Underrepresented in correspondences of TT *prý/prej* were also other expressions with the same function, namely verbs of appearance (*parecer*, *verse*) and verbs expressing intellectual states (*entender*, *suponer*). It is interesting to note that it is for these expressions with «modal overtones» that the particle *prý/prej* does not easily offer itself as an equivalent, contributing to the overall underrepresentation of the particle in the TTs.

Underrepresented are, however, most of the other correspondences of TT *prý/prej*, which indeed confirms the argument that an item without a straightforward equivalent is difficult to trigger in translation. Surprising rather than not from this perspective seems to be the fact that *que* and reporting clauses with generic subjects or reflexive verbs appear in the correspondence of *prý/prej* in both directions of translation with the same frequency. In other words, an item might be more unique in some senses than in others.

The study had to deal with methodological problems of which the fact that Czech is a small language is only one. Czech is underrepresented as source language in the parallel corpus InterCorp; in this study, our subcorpus of TT subtitles was about twice the size of the subcorpus of ST subtitles translated into Spanish, which did not allow for controlling for outliers. The subcorpus of TT fiction was about twice the size of the subcorpus of ST fiction even though only books by Spanish authors were included. Another problem concerns issues related to the quality of translation included in a parallel corpus: we noted a high frequency of *por lo visto* in the translations of *prý/prej* by one translator, which, if left unnoticed, could have skewed the data. Finally, a systematic analysis of correspondences did not identify a clear reason why the subcorpus of Cz TTs of fiction shows a predominance of cases in which the source of reported information is not present (as opposed to the known source): though we have shown that clauses and phrases introducing the source of the information (reporting clauses with referring pronouns in the subjects [quotat in Figure 7] and prepositional phrases with the preposition *según*) are underrepresented in correspondence of TT *prý/prej* as opposed to ST *prý/prej* (Figure 8), it is as well possible that what we have here is an aspect in which the texts in the two corpora are simply not comparable. This issue is, however, left for future research.

4. BIBLIOGRAPHY

AIKHENVALD, Alexandra. 2004. *Evidentiality*. Oxford: Oxford University Press.

ALTENBERG, Bengt and Sylvianne GRANGER. 2002. «Recent Trends in Cross-Linguistic Lexical Studies». In *Lexis in Contrast: Corpus-Based Approaches*, ed. by Bengt Altenberg and Sylvianne Granger. Amsterdam: John Benjamins, 3-48.

BAKER, Mona. 1993. «Corpus Linguistics and Translation Studies: Implications and Applications». In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli. Amsterdam: John Benjamins, 233-250.

CAPPELLE, Bert. 2012. «English Is Less Rich in Manner-of-Motion Verbs when Translated from French». *Across Languages and Cultures* 13 (2): 173-195. doi: 10.1556/Acr.13.2012.2.3

CATFORD, John Cunnison. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics.* London: Oxford University Press.

CHESTERMAN, Andrew. 2007. «What Is a Unique Item?». In *Doubts and Directions in Translation Studies*, ed. by Yves Gambier, Miriam Shlesinger, and Radegundis Stolze. Amsterdam: John Benjamins, 3-13.

CHLUMSKÁ, Lucie and Olga RICHTEROVÁ. 2014a. «Jak zkoumat překladovou češtinu. Výzkum simplifikace na korpusu Jerome». *Korpus, gramatika, axiologie* 9: 16-29.

CHLUMSKÁ, Lucie and Olga RICHTEROVÁ. 2014b. «Překladová čeština v korpusech». *Naše řeč* 4-5: 259-269.

ČERMÁK, František, Patrick CORNESS, and Aleš KLÉGR. 2010. *InterCorp: Exploring a Multilingual Corpus*. Prague: NLN.

GAST, Volker. 2012. «Contrastive Linguistics: Theories and Methods». In *Dictionaries of Linguistics and Communication Science: Linguistic Theory and Methodology*, ed. by Bernd Kortmann and Johannes Kabatek. Berlin: Mouton de Gruyter.

GREBEŇ, Michal. 2019. «The Type Nouns *Kind of* and *Sort of* and Their Translations Equivalents». Masters dissertation. Palacký University Olomouc.

HIRSCHOVÁ, Milada and Soňa SCHNEIDEROVÁ. 2012. «Evidenciální výrazy v českých publicistických textech (případ údajně – údajný)». In *Grammar and Corpora 2012: 4th International Conference*. Praha: Ústav pro jazyk český AV ČR – Hradec Králové: Gaudeamus.

JOHANSSON, Stig. 2007a. «Seeing through Multilingual Corpora». In *Corpus Linguistics 25. Years On*, ed. by Roberta Facchinetti. Amsterdam: Rodopi, 51-72.

JOHANSSON, Stig. 2007b. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies.* Amsterdam: John Benjamins.

KRZESZOWSKI, Tomasz P. 1990. *Contrasting Languages: The Scope of Contrastive Linguistics.* Berlin: Mouton de Gruyter.

KRUGER, Haidee. 2018. «Expanding the Third Code: Corpus-Based Studies of Constrained Communication and Language Mediation». In *Book of Abstracts: Using Corpora in Contrastive and Translation Studies Conference*, ed. by Sylviane Granger, Marie-Aude Lefer, and Lara Aguiar de Souza Penha Marion. https://cdn.uclouvain.be/groups/cms-editors-cecl/ucts2018/UCCTS2018_book_of_abstracts_01.pdf.

LEVENSTON, Edward A. 1965. «The "Translation Paradigm": A Technique for Contrastive Syntax». *International Review of Applied Linguistics* 3: 221-225.

MALÁ, Markéta. 2012. «Translation Counterparts as Markers of Meaning. The Case of Copular Verbs in a Parallel English-Czech Corpus». *Languages in Contrast* 13:170-92.

MARTINKOVÁ, Michaela. 2014. «K metodologii využití paralelních korpusů v kontrastivní lingvistice». *Naše řeč* 97 (3-4): 270-85.

MARTINKOVÁ, Michaela. 2018. «K takzvané sémantické typologii jazyků: Co česká slovesa pohybu mohou vypovídat o angličtině a španělštině». *SALi* 2018 (2): 37-53.

MARTINKOVÁ, Michaela, and Markéta JANEBOVÁ. 2017. «What English Translation Equivalents Can Reveal about the Czech "Modal" Particle *prý*: A Cross-Register Study». In *Contrastive Analysis of Discourse-Pragmatic Aspects of Linguistic Genres. Yearbook of Corpus Linguistics and Pragmatics Vol. 5*, ed. by Karin Aijmer and Diana Lewis. Cham: Springer, 63-90.

ROSEN, Alexandr, Martin VAVŘÍN, and Adrian Jan ZASINA. 2018. «*Korpus InterCorp* – čeština, verze 11 z 19.10.2018». Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz.

SELINKER, Larry. 1972. «Interlanguage». *International Review of Applied Linguistics* 10: 209-241.

ŠPÍNOVÁ, Adéla. 2018. «Hypotéza Unique Items v překladu z angličtiny». *Acta Universitatis Carolinae Philologica* 2: 117-130.

ŠTICHAUER, Pavel, and Petr ČERMÁK. 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: 7-20.

TIRKKONEN-CONDIT, Sonja. 2002. «Translationese – a Myth or an Empirical Fact?». *Target* 14 (2): 207-20.

TIRKKONEN-CONDIT, Sonja. 2004. «Unique Items – Over- or Under-Represented in Translated Language?». In *Translation Universals. Do They Exist?*, ed. by Anna Mauranen and Pekka Kujamäki. Amsterdam: John Benjamins, 177-184.