The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology

Olga Nádvorníková (Prague)

ABSTRACT

This paper analyses the frequency of six punctuation marks (the comma, period, colon, semicolon, question mark and exclamation mark) in three languages (English, French and Czech) in three different types of corpora — comparable web corpora, large monolingual general (reference) corpora and parallel (translation) corpora. The aim of the analysis is to identify which type of corpus and which methodology are the most suitable for contrastive research into punctuation. The data shows that the frequency of different punctuation marks is very sensitive to the text type. Therefore, the web corpora, containing uncontrollable amounts of various text types, cannot provide specific and reliable information about the use of punctuation marks in a given language. We argue that despite their limitations in terms of size and composition as well as the potential specific features of the language of translation, the parallel corpora used in combination with the general (reference) corpora provide the best data for such research.

KEYWORDS:

comparable web corpora, contrastive analysis, parallel corpora, punctuation, reference corpora

DOI

https://doi.org/10.14712/18059635.2020.1.2

1. INTRODUCTION

Punctuation marks fulfil multiple functions in a text — syntactic (grammatical), prosodic and semantic-pragmatic (Védénina 1980, Quirk et al. 1985, Meyer 1987, Catach 1994, Pravdová et al. 2004, Cvrček et al. 2010, Grevisse and Goosse 2011, etc.). In their syntactic function, punctuation marks separate words, clauses and sentences and mark the logical relationship between them (see Pagnoulle 2004). At the prosodic level, they indicate pauses and vary the rhythm of the text[1] while at the semantic-pragmatic level, they indicate the modality of the sentence.

Although punctuation marks share the same basic properties across many languages — they are "interlingual" (Rey-Debove 1978, Ponge 2011 and others), their frequency and specific usage may vary according to local conventions and traditions. The aim of this paper is to analyse the frequency of six punctuation marks (three intersentential: the full stop, question mark and exclamation mark, and three intra-

[1] As shown in Quirk et al. (1985) and especially in Primus (2007), the traditional assumption that there is a close association between punctuation and prosody is problematic.

sentential: the comma, colon and semicolon)[2] in three different languages (English, French and Czech) and to find out which type of corpora and which methodology are the most appropriate for contrastive research into punctuation. The analysis will be conducted on three types of corpora: comparable web corpora (Aranea), general (reference) corpora and parallel (translation) corpora. We assume that in comparison with parallel and general (reference) corpora, the advantage of the Aranea corpora will be their comparability in terms of size, composition and date of crawling (Benko 2014). However, the general (reference) and the parallel corpora benefit from the high reliability of the data and the metadata.

The paper is organised as follows: Section 2 presents the three types of corpus used in the research, which are then analysed individually in Section 3. The final section of this study summarises the results and discusses open questions for future research.

2. CORPORA

As mentioned above, three types of corpora were used in this research: comparable web corpora (Aranea, Benko 2015), large general (reference) corpora (the British National Corpus, FRANTEXT, Est républicain, and SYNv6) and a parallel (multilingual) corpus (InterCorp).

The Aranea web corpora (see Benko 2014, 2015 and 2017, http://aranea.juls.savba.sk/aranea_about/index.html) are a set of large monolingual corpora that represent 14 languages: English, including African English and Asian English, Czech, Finnish, French, German, Slovak, Spanish, Hungarian, Italian, Dutch, Polish, Portuguese, Ukrainian and Russian.[3] The corpora were created using the SpiderLing web crawler at approximately the same time (in 2013).[4] In their main variant (Maius), they have the same size of approximately 1.2 billion tokens each, contain similar (web-specific) text types, and are available via the same access platform.[5] For all these reasons, these corpora "(to a large extent) deserve the designation of being 'comparable'" (Benko 2014: 247).

General (monolingual) corpora are supposed to be representative and balanced with regard to the variety of a given language (see McEnery — Xiao — Tono 2006: 59);

[2] This list is not exhaustive and other punctuation marks could have been included in the research; in particular, ellipsis dots (suspension points), dashes (en-dashes and em-dashes, see the contrastive research conducted by Rodríguez-Castro 2011), as well as brackets and parenthesis, quotation marks, and punctuation marks operating at the level of the word (hyphens and apostrophes) or at the level of text (paragraphs, font changes, lists, etc.).

[3] Benko (2014) mentions only 11 languages; other languages have since been added (African English and Asian English, Finnish, Dutch and Portuguese).

[4] The deduplication was carried out by the Onion utility based on n-grams (Benko 2014: 250 and Benko 2013).

[5] In our research, we accessed the corpora via the corpus manager KonText of the Institute of the Czech National Corpus, based on the SketchEngine interface.

a *reference corpus* should moreover remain unchanged. The British National Corpus (BNC) is a typical example of such a corpus, as is the Czech reference (versioned) corpus SYNv6 — they both contain a large amount of texts representing the main imaginative/fiction, non-fiction and journalistic text types.[6] However, a representative reference corpus is not available for French (see Nádvorníková 2007). The corpus closest to a representative reference corpus is FRANTEXT (www.frantext.fr), a large corpus composed mostly of fiction and partly also of non-fiction (from the 12th to the 21st centuries). In our research, we used the subcorpora of FRANTEXT of fiction and non-fiction published after 1950 (31,610,109 and 18,261,370 tokens, respectively). To analyse French journalistic texts, we used the *Est républicain 2* corpus (87,984,773 tokens), again accessible via KonText. The corpus contains articles from the French regional journal, also called *L'Est Républicain*, published in 1999, 2002 and 2003.

The InterCorp parallel corpus (www.korpus.cz/intercorp) is a large multilingual corpus made up of 40 languages, with Czech as the pivot language. The whole corpus contains 2,108 billion tokens (www.korpus.cz/intercorp and Čermák and Rosen 2012 and Nádvorníková 2016 in French). It has been used extensively in research both in contrastive linguistics and in translation studies (see www.korpus.cz/biblio). The corpus is divided into a core part and collections. The core mostly consists of fiction and partly of non-fiction. The collections are composed of various text types: movie subtitles, Acquis communautaire, transcripts of debates in the European Parliament and journalistic texts (collections SYNDICATE and Presseurop). In 2017, 18 translations of the Bible were also added to the corpus. The distinction between the core and the collections is important for our research: all core texts are proof-read and alignment in this part of the corpus is manually checked and corrected in a parallel text editor (see Čermák — Rosen 2012). For this reason, the quality of the output is higher than in the collections, not proof-read and aligned only automatically, without manual checking. This difference is crucial in contrastive research, and especially in contrastive research into punctuation, very sensitive to the quality of the corpus.

3. USE OF PUNCTUATION MARKS IN ENGLISH, FRENCH AND CZECH

Several factors underlie the differences in English, French and Czech punctuation. First, there are differences in the overall principles of punctuation in the three languages. In Czech, the dominant principle governing the use of punctuation marks is the syntactic (grammatical) one, especially in the use of the comma. In English, and more so in French, the prevailing principle is more prosodic and logical. This difference can be illustrated by the use of the comma in relative clauses: in Czech, all relative clauses are separated from the rest of the sentence by commas, whereas in English and in French the presence or absence of the comma marks the distinc-

[6] In contrast to SYNv6, the BNC corpus contains a section of spoken language (approximately 10% of the whole corpus). This section was not used in our analysis, as punctuation is intrinsically related to the written language.

tion between explicative and determinative relative clauses (see Vinay and Darbelnet 1995: 189).

There are also specific conventions concerning individual punctuation marks in accordance with the stylistic and typographic traditions of the corresponding linguistic community. For example, French uses more commas than English to separate the initial adverbial from the rest of the sentence (see Guillemin-Flescher 1981: 139, Ponge 2011: 132 and Vinay and Darbelnet 1995: 189).[7] Overuse of exclamation marks is considered a potential signal of emotionalism or limited powers of self-expression in English (Newmark 1988: 58) as well as in Czech.[8] There are also important differences in the use of the semicolon, which is more frequent in French than in English and considerably more frequent in French than in Czech (see Newmark 1988: 58 for English-French and Nádvorníková and Šotolová 2016: 203–204 for French-Czech).

Finally, according to Fabricius-Hansen (1996, 1998 and 1999), the use of punctuation marks may also reflect differences in the norms that languages follow with respect to *informational density*, i.e. the way of packaging discourse information in sentences. Comparing Norwegian and German non-fictional texts on the basis of discourse representation theory (DRT), Fabricius-Hansen shows that German, preferring high informational density, encodes information in long, hierarchical sentences. In contrast, Norwegian, based on the low informational density principle, has a more paratactic, incremental style. Thus, when translating from German into Norwegian, sentences are split more frequently than in the opposite direction (see Nádvorníková 2017a for a quantitative analysis of this issue in English, French and Czech). Therefore, a high relative frequency of final punctuation marks may be inversely proportional to the informational density of a text. Fabricius-Hansen (1999: 204) adds that a more refined use of punctuation marks, including use of the colon and semicolon, also correlates with a high degree of informational density. This issue goes, however, beyond the largely quantitative scope of this study, since many other important factors influence informational density defined in this way, e.g. the text type (fiction vs non-fiction, Nádvorníková and Šotolová 2016) and the number of finite and non-finite clauses within the sentence (Fabricius-Hansen 1999).

3.1 USE OF PUNCTUATION MARKS IN THE ARANEA COMPARABLE WEB CORPORA

The advantage of the Aranea web corpora in linguistic research is the large size, comparability and a wide range of contemporary text sources, which range from advertisements for water sports equipment to amateur poetry. Another advantage is that these corpora represent different varieties of language, e.g. French spoken/written in France as well as French used in Canada, Switzerland and Belgium. Despite the recent date of compilation, the Aranea corpora have already been used in monolingual

[7] In this study, we do not deal with the purely typographic difference between the English *decimal point* (3.96) and the French *virgule décimale* (3,96).

[8] Certain researchers see the differences in the use of exclamation marks as a result of the different "mentality" of linguistic communities (Rybák 1986: 181, in a comparison of Russian and Slovak).

as well as in contrastive research (see Wachtarczyková and Garabík 2016 for Slovak and Kratochvílová and Jindrová 2017 for Spanish and Portuguese). In this section, we will explore their usability in contrastive punctuation research.

Table 1 shows the frequency of the six punctuation marks which we focus on in this study.[9]

ARANEUM (tokens)	**Anglicum Maius (1,200,023,361)**		**Francogallicum Maius 1,200,004,721**		**Bohemicum Maius[10] 1,200,000,138**		**DIN (target-reference corpus)**		
	abs. fq.	**ipm**	**abs. fq.**	**ipm**	**abs. fq.**	**ipm**	**EN–FR**	**CS–FR**	**CS–EN**
comma	44,867,269	*37,389*	50,473,489	*42,061*	59,030,867	*49,192*	–6	8	14
full stop	36,977,921	*30,814*	32,562,309	*27,135*	44,273,355	*36,894*	6	15	9
colon	3,609,504	*3,008*	5,869,354	*4,891*	5,291,201	*4,409*	–24	–5	19
question mark	2,093,370	*1,744*	2,135,119	*1,779*	2,437,335	*2,031*	–1	7	8
semicolon	1,613,443	*1,345*	1,313,170	*1,094*	488,442	*407*	10	–46	–54
exclam. mark	1,416,417	*1,180*	2,501,888	*2,085*	1,256,457	*1,047*	–28	–33	–6

TABLE 1. Frequency of six punctuation marks in English, French and Czech in the Aranea web corpora (abs. fq. = absolute frequency; ipm = instance per million, relative frequency). The highest values among the three languages are highlighted. All the differences between the three languages are statistically significant at the level of p < .001 (according to a chi-squared test).

We can see that in the three languages, the most frequent punctuation marks are the comma and full stop. This observation corresponds to common expectations and to the findings of previous research (see Quirk et al. 1985: 1613, Nádvorníková and Šotolová 2016: 203–204). First, given the above-mentioned differences between the punctuation systems in Czech in comparison with English and French, it is not surprising that the comma is most frequent in Czech. Second, the relative frequencies of the full stop corroborate the hypothesis regarding the difference between the three languages in terms of informational density: the full stop is most frequent in Czech, less frequent in English, and the least frequent in French, which is considered the densest. Finally, the frequency of exclamation marks corroborates the traditional statements concerning the differences between French, on the one hand, and Czech

[9] We searched the corpus using the following regular expressions: [word=","], [word="\."]</s> (we added the end-of-sentence attribute to eliminate the decimal point or the points after abbreviations), and [word=":"], [word="\?"], [word=";"] and [word="\!"].

[10] For Czech (and Russian and Slovak), the “Maximum” version, containing more than 3 billion tokens, is available. Nevertheless, the advantage of a bigger corpus would not compensate for the loss of comparability with the other corpora, where such a larger amount of data is not available. However, for monolingual research, the Maximum version is a valuable source of data (e.g. for research on neologisms, see Mudrochová 2019).

and English, in which this punctuation mark is encountered less frequently, on the other.[11] Nevertheless, all these results have to be verified in a twofold analysis: first, to ascertain the reliability of the composition of the corpus; second, to analyse in detail the statistical significance of the differences observed in Table 1.

We checked the composition of the corpus by first looking at the absolute and relative frequencies of the punctuation marks according to the different <doc>; i.e. the different URL pages from which the corpus is compiled. Taking, for example, the frequency of the full stop in the French Araneum corpus, we observe that the corpus is fairly diversified. The examples are taken from more than 100,000 URL pages, with an average of 305 occurrences per page. Some pages are systematically more represented than others (in absolute numbers); in the French corpus, for example, the Swiss page www.abbaye-saint-benoit.ch contains 85,136 full stops. (This page also tops the list for commas, question marks, colons and semicolons.) Nevertheless, owing to the extremely large size of the corpus, this number amounts to only 0.2% of the total number of full stops in the corpus.[12]

Another potential issue associated with web corpora is quality of data. First, it is necessary to identify the language of the URL. The overwhelming majority of the pages have been tagged in the correct language but because of the extremely large size of the corpus, mistakes are inevitable. In the French frequency list of question marks, for example, we find a web page in Russian (www.airo-xxi.ru) and in the full stop frequency list a page in English (ranked second in relative frequency). However, other language modifications may be less evident. For example, the web page www.info-turk.be provided the third-highest number of full stops while the names of the editors suggest that their native language may not be French.

More importantly, the frequency of punctuation marks may be skewed by an overrepresentation of their specific uses. This is especially true of the semicolon, which is used frequently by e-shops for listing purchasable items. Thus, 46,654 occurrences of the semicolon in the Czech Araneum corpus, i.e. 10% of the total number of tokens, come from the webpage of the publishing house Grada. If this webpage were to be removed, the relative frequency of this punctuation mark would be even lower in Table 1, with 368 instead of 407 ipm.[13] The frequency of the semicolon is also exceptional in the English sub-corpus of Aranea: in contrast to previous contrastive research (see Section 1) and the data based on reference corpora (see Section 3.2, Tables 2–4), the frequency of the semicolon is higher in English than in French. Due to the lack of more general metadata in the web corpora, it is difficult to identify the exact reason of this difference.

[11] In all three languages, exclamation marks are frequently used in personal commentaries and blogs as well as in advertising (*Only six more shopping days 'til Christmas!*).

[12] Similarly, 228,000 commas obtained from this Swiss page represent only 0.46% of the 50 million commas in the whole corpus.

[13] Another issue associated with web corpora is the quality or the lack of proof-reading. For example, on the English site www.melindadolittle.com, apostrophes were replaced by question marks in the corpus.

The second part of the analysis of the results in Table 1 deals with the statistical significance of the differences observed between the languages. According to the results from a chi-squared test, all the differences are statistically significant at the level of $p < .001$. However, this result confirms only that enough data was available for the analysis; moreover, with such high absolute frequencies, the results are usually statistically significant. Therefore, we also tested the *effect size* of the differences by the means of the *difference index* (DIN, see Cvrček and Fidler 2015), as shown in the last three columns of Table 1 (the target corpus is always the first mentioned). The DIN takes into account only the relative frequencies in the target corpus (A) and in the reference corpus (B):

$$DIN = 100 \times \frac{relFQ(A) - relFQ(B)}{relFQ(A) + relFQ(B)}$$

The DIN values may vary between 100 and –100, 100 meaning that the analysed item occurs only in the analysed (target) corpus, and not at all in the reference corpus, and 100 meaning that the item was found only in the reference corpus. A value of 0 indicates that the relative frequency of the item is the same in both corpora. The results potentially interesting for analysis start at approximately 50, but with our data based on high absolute frequencies, the threshold of significance may be even lower. Considering the data presented in Table 1, we can state that, in general, the differences between the three languages in the frequency of use of punctuation marks are not considerable in the Aranea web corpora. Even the traditionally mentioned discrepancies in the use of exclamation marks do not reach the critical value threshold of 50, although they do confirm the more infrequent use of exclamation marks in Czech and English in comparison with French. The only clearly distinguished difference between the three languages is the low frequency of the semicolon in Czech in comparison with English and French.

In order to identify the potential idiosyncrasies of the web corpora, we will now compare this data with that obtained from the large monolingual representative general corpora of the three languages.

3.2 USE OF PUNCTUATION MARKS IN MONOLINGUAL GENERAL CORPORA

Monolingual general corpora, such as the British National Corpus or the SYN corpus of the Czech National Corpus, are designed to represent as much as possible a given language or language variety as a whole (see McEnery — Xiao — Tono 2006: 15). An advantage of these corpora is the reliability of the data, since the texts, chosen carefully according to a specific sampling frame, are supplemented with detailed metadata. Thanks to these metadata, it is possible to select subcorpora corresponding to specific text types, which most often are fiction, non-fiction and journalistic. Furthermore, the size of the general corpora is large enough even for very advanced research.

However, scholars conducting contrastive research have to deal with the fact that these corpora are not comparable for different languages, neither in size (cf. Tables 2, 3 and 4), nor in composition. In non-fiction, for example, the three corpora contain — in different proportions — texts from various scientific domains. These are academic as well as non-academic and texts ranging from quantum physics to texts about sociology and the automobile industry. Texts classified as "journalistic" or "newspaper" are also heterogeneous: tabloids, sports news, editorials, commentaries, etc., use very different language — and different punctuation. Finally, the category of "imaginative" texts may include poetry, drama, novels, etc. In order to assure (relative) comparability with the core of the parallel corpus used in Section 3.3, we restricted this last text type to novels, i.e. "prose" in the BNC and in SYNv6 and "romans" in FRANTEXT.[14]

Tables 2, 3 and 4 show the absolute and relative frequencies of the six analysed punctuation marks in the three general monolingual corpora in fiction, non-fiction and journalistic text types, including the difference index (DIN) for this data in comparison with data from the Aranea corpora (see Table 1). In the comparison, the Aranea corpora are the target corpora and the general monolingual corpora are the reference corpora. The DIN values show that the frequencies of punctuation marks observed in the web corpora (Table 1) are different from results obtained in the monolingual reference corpora. All the differences between the Aranea corpora and the corresponding monolingual corpora are statistically significant (according to a chi-squared test) and the highest values (DIN and ipm) among the three languages are highlighted.

Fiction	**BNC (15,644,928)**		**DIN (Aranea)**	**FRANTEXT (31,610,109)**		**DIN (Aranea)**	**SYNv6 (22,493,302)**		**DIN (Aranea)**
	abs. fq.	**ipm**		**abs. fq.**	**ipm**		**abs. fq.**	**ipm**	
comma	973,841	*62,246*	–25	2,222,988	*70,325*	–25	1,802,248	*80,124*	–24
full stop	929,584	*59,418*	–32	1,354,006	*42,835*	–22	1,155,414	*51,367*	–16
colon	18,611	*1,190*	43	130,408	*4,126*	8	75,709	*3,366*	13
question mark	138,914	*8,879*	–67	149,258	*4,722*	–45	152,402	*6,775*	–54
semicolon	27,806	*1,777*	–14	61,091	*1,933*	–28	18,538	*824*	–34
exclam. mark	46,518	*2,973*	–43	124,591	*3,941*	–31	118,051	*5,248*	–67

TABLE 2. Frequency of six punctuation marks in English, French and Czech in fiction in general corpora (abs.fq = absolute frequency; ipm = instance per million, relative frequency). The DIN shows the difference between the general (reference) corpus and the corresponding Aranea corpus (target corpus).

[14] To make sure that our corpora are representative of contemporary language, we used only those sub-corpora containing texts published after 1950. Still, the corpora are not fully comparable: texts (samples) in the BNC were published mainly between 1975 and 1993, the majority of texts in the fiction and non-fiction subcorpus in FRANTEXT were published between 1950 and 1980, whereas the *SYNv6* subcorpora (and *Est républicain*) contain texts published mainly in the 21st century. In all the sub-corpora, only non-translated (original) texts were used in this research. The BNC and FRANTEXT do not contain translations; in SYNv6, the source language was limited to Czech.

Non-fiction	BNC (31,885,063)		DIN (Aranea)	FRANTEXT (18,261,370)		DIN (Aranea)	SYNv615 (36,618,042)		DIN (Aranea)
	abs. fq.	ipm		abs. fq.	ipm		abs. fq.	ipm	
comma	1,513,828	47,478	−12	1,126,234	61,673	−19	2,382,766	65,071	−14
full stop	1,155,912	36,252	−8	564,010	30,885	−6	1,466,519	40,049	−4
colon	78,418	2,459	10	86,576	4,741	2	124,123	3,390	13
question mark	27,868	874	33	23,004	1,260	17	37,284	1,018	33
semicolon	91,031	2,855	−36	72,088	3,948	−57	57,448	1,569	−59
exclam. mark	5,490	172	75	9,703	531	59	18,748	512	34

TABLE 3. Frequency of six punctuation marks in English, French and Czech in non-fiction in general corpora (abs.fq = absolute frequency; ipm = instance per million, relative frequency). The DIN shows the difference between the general (reference) corpus and the corresponding Aranea corpus (target corpus).

Journalistic	BNC (10,527,721)		DIN (Aranea)	Est républicain (87,984,773)		DIN (Aranea)	SYNv6 (214,186,383)		DIN (Aranea)
	abs. fq.	ipm		abs. fq.	ipm		abs. fq.	ipm	
comma	442,385	42,021	−6	5,134,134	58,353	−16	12,560,703	58,644	−9
full stop	404,819	38,453	−11	3,307,902	37,596	−16	11,592,891	54,125	−19
colon	42,306	4,019	−14	519,973	5,910	−9	465,457	2,173	34
question mark	7,787	740	40	71,800	816	37	549,892	2,567	−12
semicolon	10,687	1,015	14	326,919	3,716	−54	26,280	123	54
exclam. mark	2,503	238	66	85,479	972	36	177,322	828	12

TABLE 4. Frequency of six punctuation marks in English, French and Czech in journalistic texts in general corpora (abs.fq = absolute frequency; ipm = instance per million, relative frequency). The DIN shows the difference between the general (reference) corpus and the corresponding Aranea corpus (target corpus).

The comparison of the data obtained from the general corpora, independently from the web corpora, confirmed some of the traditional contrastive observations mentioned in Section 2. For example, we can observe a higher relative frequency of the comma and the full stop in Czech, with the exception of the full stop in fiction, where the relative frequency is higher in English than in Czech. The data also shows a systematically high relative frequency of the colon and semicolon in French, indicating a more refined, i.e. more elaborated use of punctuation mentioned by Fabricius-

[15] Non-fiction in SYNv6 was limited to books from the categories PRO, SCI and POP. In FRANTEXT, this genre includes "essais" and "traités". In the BNC, the domain "informative" includes academic and non-academic books and periodicals.

Hansen (1999: 204; see Section 1). Nevertheless, because of the limited comparability[16] of the different corpora, no firm conclusions can be drawn.

According to the DIN values, the results from the Aranea web corpora are closest to those from non-fiction and journalistic texts in the reference corpora, especially with regard to the frequency of the comma and full stop, and partly the colon. Figure 1, which summarises the DIN values for the three languages and the three text types, shows this tendency more clearly:

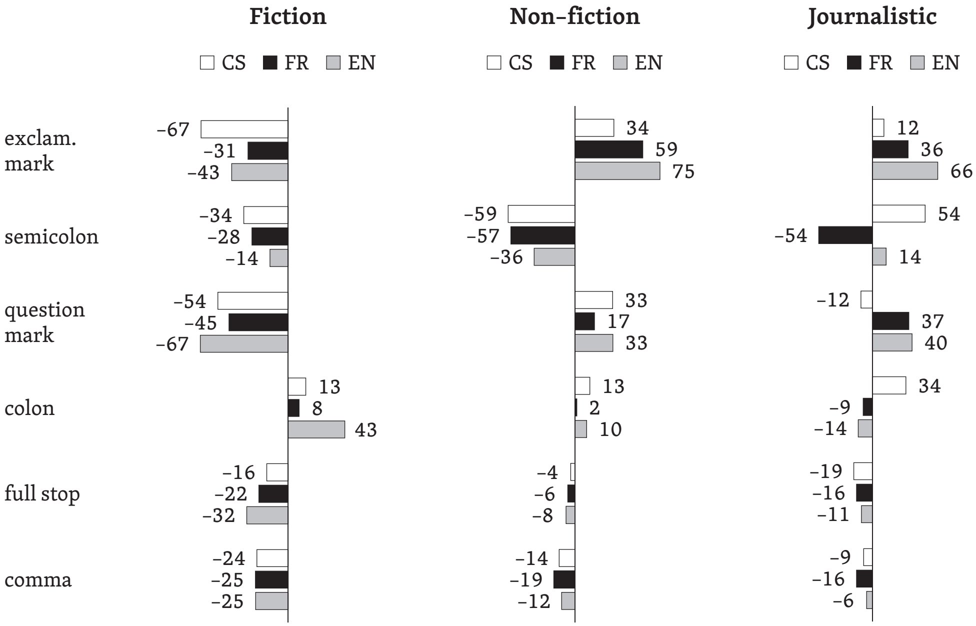

FIGURE 1. DIN values comparing the relative frequency of six punctuation marks in general (reference) corpora and in the Aranea web corpora in Czech, French and English in the fiction, non-fiction and journalistic text types. The target (analysed) corpus is Aranea; the reference corpus is general (monolingual) corpus.

Although the DIN values between the Aranea web corpora and the general (reference) corpora in the frequency of the comma, full stop and partly the colon indicate the similarity of web corpora and non-fiction/journalistic texts, such a tendency is not observed for the other punctuation marks. In fact, the differences in frequency of the other punctuation marks seem to be more sensitive to the text type of the reference corpus than to the difference in language. For example, question marks and exclamation marks are significantly more frequent in fiction in the three languages

[16] For example, from the very high relative frequency of the question mark in Czech in Table 4, we cannot conclude that Czech journalists ask more questions than their English or French colleagues.

than in the web corpora because of dialogues typical for fiction. Conversely, in non-fiction, the exclamation mark is significantly less frequent than in the web corpora.[17]

This tendency is also clear from the comparison of the individual general corpora. As shown in Figure 2, based on the data from Tables 2–4, more important differences in the frequency of punctuation marks may be observed *not* between the different languages within the same fiction text type, but between the different text types *within* the same language (cf. the same observation, on different data, in Kruger & Van Rooy 2018 and in Chlumská 2017):

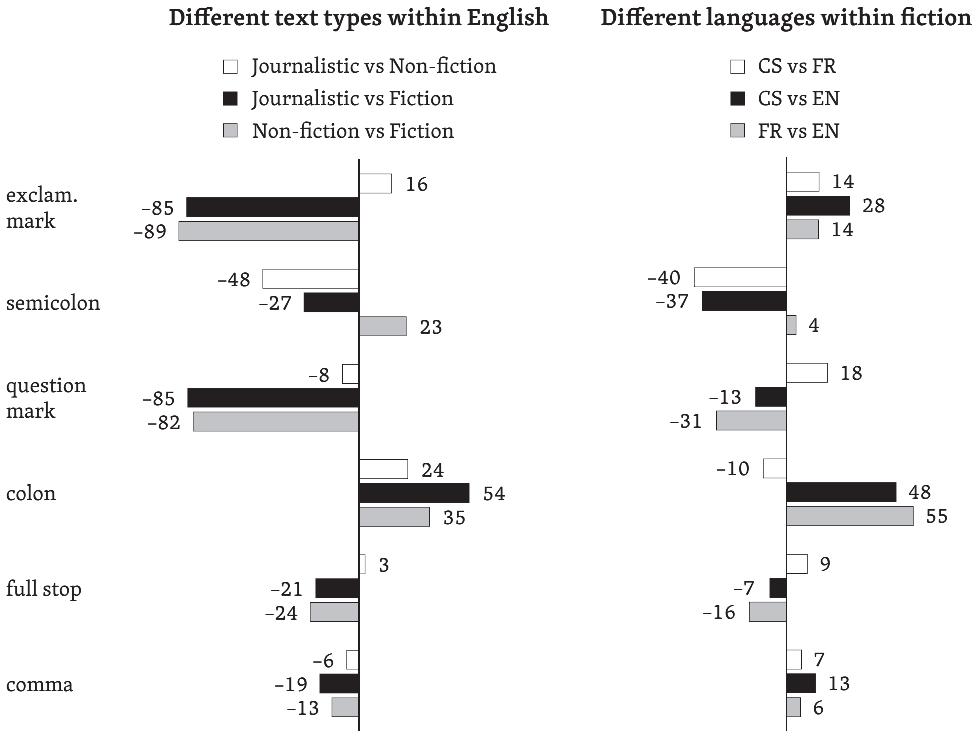

FIGURE 2. DIN values comparing the relative frequency of six punctuation marks in general (reference) corpora in the fiction, non-fiction and journalistic text types in English (left side) and in English, French and Czech within one text type — fiction (right side). The analysed (target) corpus is always the first mentioned.

[17] The high frequency of the exclamation mark in Czech fiction, which surprisingly is higher than in French, may be caused by the presence of some children's literature texts in the Czech sub-corpus that make up 9% of the entire corpus. Nevertheless, after a closer look at the data, we can conclude that only 19% of exclamation marks come from children's literature. For the corpus research of the specific features of children's literature and its translation, see Čermáková and Chlumská (2016) or Čermáková (2017).

This tendency to more important differences between different text types within the same language than between different languages within the same text type (see Figure 2) is crucial for the methodology of contrastive research into punctuation: it is necessary to specify as precisely as possible the text type of the corpora compared in the different languages. For this reason, we limited the corpus in the last section of this research, devoted to parallel (translation) corpora, not only to fiction, but more specifically to novels.

3.3 USE OF PUNCTUATION MARKS IN PARALLEL CORPORA

Unlike the types of corpora in Sections 3.1 and 3.2, parallel (translation) corpora are neither large in size, nor comparable (unlike the Aranea corpora). Their main advantage for contrastive research is that they allow for a direct comparison between the source sentence (source text) and its translation(s).

In this study, we limited the InterCorp parallel corpus to fiction available in the three languages. For this reason, although the size of bidirectional subcorpora, for example only French-Czech or English-Czech, may be quite important (18,953,496 tokens in 165 texts for the English-Czech subcorpus), the intersections of the three languages, used in this research, are much less considerable: 1,134,556 tokens (11 texts by 6 authors) for translations from English into Czech and French, 1,483,802 tokens (19 texts by 6 authors) for translations from Czech, and only 573,088 tokens (6 texts by 6 authors) for translations from French into Czech and English. The limited size of the subcorpora increases the sensitivity of the data to the composition of the corpus. This is particularly poignant in translations from English, problematic due to the over-representation of texts written by one author (half of the corpus — 590,313 tokens — come from four novels by J. K. Rowling) and because several texts were written in the 19th century (the authors are Rudyard Kipling, H. G. Wells and Lewis Carroll).

Table 5 shows the absolute and relative frequencies of the six punctuation marks in original texts (EN, FR, CS) and in the corresponding translations (e.g. en(FR) means English translation from French).

As we did for the Aranea corpora, we can first verify the reliability of the data by comparing the relative frequencies in parallel corpora with those obtained in the reference corpora (see Table 2 — fiction). With regard to the original (non-translated) texts, we can conclude that most of the differences do not go beyond DIN 20 (see Table 5). Therefore, the corpus is quite reliable for the chosen text type. The differences concerning the original (non-translated) texts are due to the above-mentioned limited size and specific composition of the sub-corpora. For example, in English, the higher frequency of the exclamation mark in InterCorp, in comparison with its frequency in the BNC (DIN value 30), is associated with the influence of four Harry Potter novels by J. K. Rowling (there are 3,426 exclamation marks in the Harry Potter subcorpus, which constitutes more than a half of the whole number of occurrences of this punctuation mark in English originals).[18] Similarly, a slightly

[18] The relative frequency of exclamation marks is lower in the first Harry Potter books than in the later ones (cf. 4,783 ipm in *The Philosopher's Stone* vs. 7,399 ipm in *The Prisoner of Azkaban*).

InterCorp	**size of the corpus**	**comma**		**full stop**		**colon**		**question mark**		**semicolon**		**exclam. mark**	
		abs.fq	**rel.fq**	**abs.fq**	**rel.fq**	**abs.fq**	**rel.fq**	**abs.fq**	**rel.fq**	**abs.fq**	**rel.fq**	**abs.fq**	**rel.fq**
EN (orig.)	1,134,556	65,338	*57,589*	47,498	*41,865*	1,589	*1,401*	**6,037**	***5,321***	**3,634**	***3,203***	6,201	*5,466*
DIN (BNC)			−4		−17		8		−25		29		30
fr(EN)	1,210,677	68,314	*56,426*	**54,642**	*45,133*	**2,507**	*2,071*	5,881	*5,184*	1,088	*899*	**6,573**	*5,429*
DIN FRANTEXT			−11		3		−33		5		−37		16
cs(EN)	1,014,054	**77,089**	***76,021***	48,097	***47,430***	2,163	***2,133***	5,926	*5,223*	3,047	*3,005*	6,354	***6,266***
DIN (SYNv6)			−3		−4		−22		−13		57		9
FR (orig.)	573,088	**36,363**	***63,451***	22,956	*40,057*	1,322	*2,307*	2,009	*3,506*	1,303	*2,274*	**3,383**	*5,903*
DIN FRANTEXT			−5		−3		−28		−15		5		20
en(FR)	548,054	31,886	*58,180*	**22,991**	*41,950*	734	*1,339*	**2,025**	*3,695*	**1,980**	*3,613*	2,691	*4,910*
DIN (BNC)			−3		−17		6		−41		34		25
cs(FR)	471,045	22,749	*48,295*	22,749	***48,295***	**1,405**	***2,983***	1,967	***4,176***	1,976	***4,195***	2,924	***6,207***
DIN (SYNv6)			−25		−3		−6		−24		67		8
CS (orig.)	1,483,802	**104,634**	***70,517***	58,914	***39,705***	5,562	***3,748***	6,639	***4,474***	3,412	*2,299*	4,459	***3,005***
DIN (SYNv6)			−6		−13		5		−20		47		−27
en(CS)	1,671,248	81,792	*48,941*	**65,425**	*39,147*	4,922	*2,945*	**7,193**	*4,304*	**3,417**	*2,045*	4,504	*2,695*
DIN (BNC)			−12		−34		42		−35		7		−5
fr(CS)	1,735,949	97,001	*55,878*	62,114	*35,781*	**6,452**	*3,717*	6,969	*4,015*	4,375	***2,520***	**5,036**	*2,901*
DIN FRANTEXT			−11		−9		−5		−8		13		−15

TABLE 5. Frequency of six punctuation marks in English, French and Czech in the parallel (translation) corpus InterCorp (abs.fq = absolute frequency; ipm = instance per million, relative frequency). The values of DIN higher than 30 are highlighted. The DIN shows the difference between the parallel sub-corpus and the corresponding general sub-corpus for fiction (Table 2). All the differences are statistically significant according to a chi-squared test.

lower frequency of the full stop in the Czech part of the InterCorp parallel corpus, in comparison with its frequency in SYNv6, is influenced by the presence of two novels by Bohumil Hrabal, known for his long sentences giving the impression of a free flux of narration. Some differences, however, are less obvious: for example, in the French parallel subcorpus, the relative frequency of the semicolon is similar to its frequency in the reference (general) corpus, alhough 1,076 semicolons out of 1,303 (83%) come from the same text: *Les particules élémentaires* by Michel Houellebecq. Thus, the analysis of the use of the semicolon in this sub-corpus may be skewed by the specific idiolect of one author.

With regard to the translations, the two potential approaches to their assessment reveal the translator's paradox. We either expect to observe similarity in the use and the frequency of punctuation marks in translations with their use and frequency in the reference corpora, which indicates adherence to target language punctuation conventions (cf. the DIN values in Table 5), or we expect to observe similarity with the source text(s), indicating adherence to the style of the author of the source text. Failure to follow target language conventions may be caused by interference from the source language, which may create the effect of "translationese" (see Bystrova-McIntyre 2007, Øverås 1998 or Rogríguez-Castro 2011, etc.). However, non-motivated changes perpetrated by the translator may entail simplification, explicitation and normalisation of the source text (see the abundant literature on so-called "translation universals", Baker 1993; Mauranen and Kujamäki 2004; Pápai 2004; Malmkjær and Windle 2012 or Robin 2017).

The translator's task is even more difficult for literary texts, in which the use of punctuation is often creative and may violate source language conventions (see also Primus 2007: 43). Ponge (2011: 133) points out that the translator has to be able to distinguish the inherent differences between the source and target languages as well as original, individual choices of the author, which should be conveyed in the target text. Several papers in translation studies have shown that normalisation, explicitation and simplification of the creative use of punctuation in literary translation may modify or even erase the original style of the source text. This can be seen in the Russian and French translations of novels by William Faulkner and Virginia Woolf (cf. May 1997), the English translations of Dutch novels (Vanderauwera 1985) and of Hans Christian Andersen's fairy tales (cf. Malmkjær 1997), a Spanish translation of Marcel Proust's *A la recherche du temps perdu* (cf. Ponge 2011), or several translations from French into Czech and especially from Czech into French (cf. Nádvorníková and Šotolová 2016 and Šotolová 2013).

A detailed analysis of all the contrastive data in Table 5 is beyond the scope of this study. Therefore, we will illustrate the potential use of parallel corpora in contrastive punctuation research by only analysing the semicolon. As we have seen in Sections 3.1 and 3.2, the semicolon showed the most frequent and the most considerable differences, especially between Czech, on the one hand, and English and French, on the other (cf. Tables 1–4). Table 5 shows an unexpectedly high incidence of the semicolon in the InterCorp parallel sub-corpus of Czech original fiction: in comparison with the general reference corpus (Table 2), its relative frequency is nearly three times higher (2,299 ipm in InterCorp vs. 824 ipm in SYNv6). This is mostly due to the influence of two Czech authors known for their refined use of punctuation, whose novels are included in InterCorp: Karel Čapek (*War with the Newts*) and Milan Kundera (*The Joke*, *Laughable Loves*, *The Unbearable Lightness of Being*, *Farewell Waltz* and *Immortality*).[19] However, other authors in this sub-corpus (e.g. Bohumil Hrabal and Jaroslav Hašek, author of *The Fateful Adventures of the Good Soldier Svejk During the World War*) use the semicolon rarely or not at all, as they try to emulate the spoken language.[20]

While the semicolon is used very frequently in the sub-corpus of Czech originals, its frequency in French translations of these Czech texts is even higher because French translators used semicolons in texts where the Czech authors did not: there are 0 and 4 semicolons in the Czech originals vs. 250 and 129 in the French translations of the two novels by Bohumil Hrabal, and 6 semicolons in the Czech original vs. 354 semicolons in the French translation of the novel by Jaroslav Hašek. (English translators adhered to the original style in these texts more than the French translators, with scores of 0:39 and 3:65, respectively.)[21] The use of semicolons renders the target text smoother, more logical, more structured, more "French", but erases the originality of the source text. Moreover, as we observed in Nádvorníková and Šotolová (2016), the greater use of one explicitating punctuation mark is usually accompanied by other shifts. These include explicitation by other punctuation marks, especially the colon, and by connectives, normalising the length of sentences (splitting long ones and joining short ones), etc. If explicitation and normalisation become the translator's strategy, the style of the whole source text is necessarily altered (see Nádvorníková 2017a and Nádvorníková forthcoming).

On the contrary, in the opposite direction of translation, from French into Czech, we would expect a decrease in the frequency of the semicolon, in accordance with the differences in the conventions observed in Table 2. However, the frequency of the semicolon in texts translated into Czech is surprisingly even higher than in the French source sub-corpus. To explain this result, it is necessary for us to look closer at the correspondences. Czech translators often respect the use of the semicolon in the French source text, even if this means that its frequency is higher than in Czech non-translated texts.[22] They also add semicolons to structure long target sentences containing a high number of finite clauses, frequently corresponding to complex source sentences hierarchically structured by non-finite clauses in French. This sug-

[19] These two authors are also most frequent users of the semicolon in SYNv6 (Table 2).

[20] The necessity to examine in detail the composition of the corpus is also illustrated by the sub-corpus of translations from English into French: the overall decrease in the frequency of the semicolon is due to the strategy adopted by a sole translator (J.-Fr. Ménard), who erased nearly all the original semicolons, replacing them mostly with full stops.

[21] To be absolutely honest, it is not possible to say, on the basis of this rough quantitative data, that the translator added semicolons. S/he may have used semicolons where the original had other punctuation marks, but other punctuation marks where the original had semicolons. Nevertheless, the resulting absolute frequency remains the same.

[22] Chlumská (2017: 69) also observed an increase in the frequency of the semicolon in translated texts (on the basis of the Czech comparable translation corpus *Jerome*, Chlumská 2013), especially in translations from English.

gests that the research of shifts in punctuation has to be intrinsically linked to the contrastive analysis of the syntax of the two or three languages under examination, especially with regard to informational density (see Section 1).

From a methodological point of view, it is important to point out that in this analysis of parallel data, we compared the absolute frequencies of punctuation marks in the analysed sub-corpora because the source and target texts are considered *equivalent*. In fact, the data in Table 5 shows that the relative frequencies are not 100% reliable for comparing typologically different languages. Czech, a synthetic language, encodes the same information in fewer words than relatively analytic languages, such as English and French. As the relative frequencies are based on the number of tokens, it is inevitable that the relative frequencies of punctuation marks will be higher in texts written in languages that encode information in fewer words. This can be seen, for example, in the frequency of the full stop in translations from English in Table 5: the absolute frequency is higher in translations into French than into Czech, but the *relative* frequency is higher in Czech than in French, due to the lower number of tokens in the Czech sub-corpus (1,210,677 in French vs. 1,104,054 in Czech). This observation is crucial for interpreting the data obtained from comparable web corpora and general corpora in Sections 3.1 and 3.2. In fact, all the relative frequencies for Czech should be reconsidered by a quotient corresponding to this difference, as all the relative frequencies in Czech would be lower.

Thus, parallel corpora seem to be more suitable for contrastive punctuation research than comparable or general corpora. Despite the issues caused by their limited size and by potential idiosyncrasies introduced by the translator, parallel corpora allow for more precise quantitative research due to the direct comparison of absolute frequencies. In addition, they allow for a fine-grained qualitative analysis of the shifts and correspondences of punctuation marks in individual sentences. However, general monolingual corpora are invaluable in their role as reference corpora as they reveal potential biases of both translated and non-translated texts contained in the parallel corpus (cf. a similar methodology suggested in Johansson 1998 and 2007).

4. CONCLUSION

The aim of this paper was to indentify which type of corpus and which methodology are the most appropriate for contrastive research into punctuation. The analysis based on the comparison of English, French and Czech large general monolingual corpora has shown that the frequency of punctuation marks is very sensitive to the text type (fiction, non-fiction and journalistic). For this reason, Aranea web corpora are not a suitable source of data for contrastive research into punctuation, despite their comparability in size, date of compilation and methods by which they were crawled. In fact, the exact composition of the web corpora, in terms of proportions of different text types, cannot be identified, and sources ranging from excerpts from the Bible to real estate advertisements are too heterogeneous to provide specific data about the use of punctuation marks in a given language. Moreover, apart from the text type, other uncontrollable variables have a considerable influence on the resulting frequencies in Aranea web corpora, especially the overrepresentation of various specific uses of punctuation marks, such as the use of the semicolon to separate purchasable items in e-shops.

In contrast to the web corpora, the advantage of the general monolingual corpora is the quality of metadata that allows us to specify the composition of the corpus, for example, the specific domains of non-fiction. However, their usability for contrastive research into punctuation is limited because, strictly speaking, they are not comparable in size or composition. Moreover, general monolingual corpora have to compare only relative frequencies in the languages under study, but those are not reliable when comparing typologically different languages. As revealed by the comparison of absolute and relative frequencies of punctuation marks in parallel corpora, relative frequencies are biased when data for more synthetic languages, such as Czech, was compared with data for more analytic languages, such as French and English. In fact, as more synthetic languages encode the same information in fewer words than less synthetic languages, the difference in the number of tokens of the corpus alters the relative frequency. For this reason, all the relative frequencies for Czech compared with English and French should be reconsidered by a quotient corresponding to this difference.

Thus, the advantage of parallel corpora in comparison with general monolingual corpora and comparable web corpora is the possibility to compare directly the absolute frequencies of different punctuation marks in the source texts and those in the corresponding translations, as they are considered equivalent. Parallel concordances also allow for a qualitative analysis of the shifts and correspondences of the punctuation systems in different languages. Nevertheless, several methodological limitations of parallel corpora have to be taken into account, namely their limited size and composition, specific features of the language of translation, idiolects of authors or translators, etc.

In order to avoid or at least identify the potential influence of the specific features of the language of translation, some methodological principles have to be respected while working on parallel corpora (see also Nádvorníková 2017b and 2017c). The analysis has to be bidirectional and the data for both translated and non-translated language has to be compared to the data in the reference (monolingual) corpora (see Johansson 1998). In the specific research into punctuation, the data showed that punctuation marks should be analysed not separately but as a system because a decrease in the frequency of one punctuation mark in translation is often compensated by an increase in the frequency of another punctuation mark. For example, the use of the semicolon is closely linked to the use of the comma and the full stop.

Future contrastive research into punctuation marks may focus more on qualitative analysis, considering the changes in their different uses rather than the overall frequencies. From the contrastive point of view, it will be interesting to examine the synergy of changes to punctuation and other modifications in translation, such as adding connectives or moving non-finite verb forms to finite ones, in order to identify more general structural differences between languages. Finally, but just as important, a more thorough comparison of punctuation marks in different text types such as fiction, non-fiction and journalistic and their translations may reveal specific conventions or the impact of editorial guidelines. Thus, further research into punctuation, still underexplored within linguistic research, may become a key to unlocking very complex systems.

REFERENCES

Baker, M. (1993) Corpus Linguistics and Translation Studies. Implications and Applications. In: Baker M., G. Francis and E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, 233–250. Amsterdam/Philadelphia: John Benjamins.

Benko, V. (2013) Data Deduplication in Slovak Corpora. In: *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*, 27–39. Lüdenscheid: RAM-Verlag.

Benko, V. (2014) Aranea: Yet Another Family of (Comparable) Web Corpora. In: Sojka, P., A. Horák, I. Kopeček, and K. Pala (eds) *TSD 2014*, LNAI 8655, 257–264. Springer International Publishing.

Benko, V. (2017) Are Web Corpora Inferior? The Case of Czech and Slovak. In: Bański, P. et al. (eds) *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and* NLP, 43–48. Birmingham, July 2017.

Bystrova-McIntyre, T. (2007) Looking at the overlooked: A corpora study of punctuation use in Russian and English. *Translation and Interpreting Studies* 2(1), 137–162.

Catach, N. (1994) *La Ponctuation*. Paris: PUF.

Čermák, F. and A. Rosen (2012) The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3), 411–427.

Čermáková, A. (2017) Translating children's literature: Some insights from corpus stylistics. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies* 71 (1) (special issue *The uses of parallel corpora in the stylistic analysis of films and literature for children*, ed. by M. Toolan), 117–134.

Čermáková, A. and L. Chlumská (2016) Jazyk dětské literatury: kontrastivní srovnání angličtiny a češtiny [Language of children's literature: an English and Czech contrastive study]. In: A. Čermáková, L. Chlumská and M. Malá (eds) *Jazykové paralely*, 162–187. Praha: NLN.

Chlumská, L. (2017) *Překladová čeština a její charakteristiky*. Praha: NLN.

Cvrček, V. *et al.* (2010) *Mluvnice současné češtiny. Jak se píše a jak se mluví I*. Praha: Karolinum.

Cvrček, V. and M. Fidler (2015) A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics* 23(2), 197–239.

Fabricius-Hansen, C. (1996) Informational Density: A Problem for Translation and Translation Theory. *Linguistics* 34, 521–565.

Fabricius-Hansen, C. (1998) Informational density and translation, with special reference to German — Norwegian — English. *Language and Computers* 24, 197–234.

Fabricius-Hansen, C. (1999) Information Packaging and Translation: Aspects of Translational Sentence Splitting (German — English/Norwegian) In: Doherty, M. (ed.) *Sprach-spezifische Aspekte der Informationsverteilung*, 175–214. Berlin: Akademie Verlag.

Grevisse, M. and A. Goosse (2011) *Le Bon usage*. Paris et Louvain-La-Neuve: Duculot.

Guillemin-Flescher, J. (1981) *Syntaxe comparée du français et de l'anglais: problèmes de traduction*. Paris: OPHRYS.

Johansson, S. (1998) On the role of corpora in cross-linguistic research. In: Johansson, S. and S. Oksefjell (eds) *Corpora and Cross-linguistic Research*, 3–24. Amsterdam — Atlanta: Rodopi.

Johansson, S. (2007) *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Kratochvílová, D. and J. Jindrová (2017) Ingressive verbal periphrases in Spanish and Portuguese. *Linguistica Pragensia* 27(1), 38–56.

Kruger, H. & Van Rooy, B. (2018) Register variation in written contact varieties of English: a multidimensional analysis. *English World-Wide* 39(2), 214–242.

Malmkjær, K. (1997) Punctuation in Hans Christian Andersen's stories and in their translations into English. In: Payotas, F. (ed.) *Nonverbal Communication and Translation. New Perspectives and Challenges in Literature, Interpretation and the Media*, 151–162. Amsterdam: John Benjamins.

Malmkjær, K. and K. Windle (eds) (2012) *The Oxford Handbook of Translation Studies.* Oxford: Oxford University Press.

Mauranen, A. and P. Kujamäki (ed.) (2004) *Translation Universals: Do they exist?* Amsterdam/Philadelphia: John Benjamins.

May, R. (1997) Sensible Elocution: How Translation Works in and upon Punctuation. *The Translator* 3(1), 1–20.

McEnery, T., R. Xiao, and Y. Tono (2006) *Corpus-based language studies: an advanced resource book*. London: Routledge.

Meyer, C. F. (1987) *A linguistic study of American punctuation.* Frankfurt: Peter Lang.

Mudrochová, R. (2019) La productivité et la fréquence d'emploi des verbes d'origine anglaise récemment lexicalisés dans les contextes français, québécois et tchèque. *Xlinguae* 1XL/2019, 96–108.

Nádvorníková, O. (2007) Existuje pro francouzštinu ekvivalent Českého národního korpusu? In: Štícha, Fr. and J. Šimandl (eds) *Gramatika a korpus*, 179–190. Praha: ÚJČ AV ČR.

Nádvorníková, O. (2016) Le corpus multilingue InterCorp et les possibilités de son exploitation. In: Buchi É., J. P. Chauveau and J.-M. Pierrel (eds) *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15–20 juillet 2013)*, 223–237. Société de linguistique romane/ÉLiPhi, Strasbourg. Available at http://www.atilf.fr/cilpr2013/actes/section-16.html. [last accessed 22 May 2018].

Nádvorníková, O. (2017a) 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: Emonds, J. and M. Janebová (eds) *Language Use and Linguistic Structure*, 445–461. Olomouc: Palacký University Olomouc. Available at http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf [last accessed 22 May 2018].

Nádvorníková, O. (2017b) Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela — cognition, représentation, langages* 15(1). Available at https://journals.openedition.org/corela/4810 [last accessed 22 May 2018].

Nádvorníková, O. (2017c) Le corpus multilingue InterCorp : nouveaux paradigmes de recherche en linguistique contrastive et en traductologie. *Studii de Lingvistica* 7, 67–88. Available at http://studiidelingvistica.uoradea.ro/docs/7-2017/pdf_uri/Nadvornikova.pdf [last accessed 22 May 2018].

Nádvorníková, O. (forthcoming) Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). *Research in Language*.

Nádvorníková, O. and J. Šotolová (2016) Změny v segmentaci na věty v překladových textech: analýza dat z francouzsko-českého paralelního korpusu. [Changes of Segmentation in Phrases in Translation]. In: Čermáková A., L. Chlumská and M. Malá (eds) *Jazykové paralely*, 188–235. Praha: ÚČNK/NLN.

Newmark, P. (1988) *A Textbook of Translation*. Singapore: Prentice Hall.

Øverås, L. (1998) In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta: Tranlator's Journal* 43(4), 557–570.

Pagnoulle, Ch. (2004) Traduire les points et les virgules. In: Ballard, M. and L. Hewson (eds) *Correct/Incorrect*, 33–40. Arras: Artois Presses Université.

Pápai, V. (2004) Explicitation. In: Mauranen, A. and P. Kujamäki (eds) *Translation Universals: Do they exist?*, 143–165. Amsterdam/Philadelphia: John Benjamins.

Ponge, M. (2011) Pertinence linguistique de la ponctuation et traduction (français-espagnol). *La linguistique* 47(2), 121–136.

Pravdová, M. et al. (2004) *Akademická příručka českého jazyka*. Praha: Academia.

Primus, B. (2007) The typological and historical variation of punctuation systems: Comma constraints. *Written Language and Literacy* 10(2), 103–128.

Quirk, R. et al. (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Rey-Debove, J. (1978) *Le Métalangage. Étude linguistique du discours sur le langage*. Paris: Le Robert.

Robin, E. (2017) Translation Universals Revisited. *Forum* 15(1), 51–66.

Rodríguez-Castro, M. (2011) Translationese and punctuation: An empirical study of translated and nontranslated international newspaper articles (English and Spanish). *Translation and Interpreting Studies* 6(1), 40–61.

Rybák, J. (1986) Interferencia interpunkcie (z úvah o prekladaní). *Zborník pedagogickej fakulty v Prešove* 20(3), 178–196.

Šotolová, J. (2013) Sur le point-virgule et autres détails éphémeres. *Études Romanes de Brno* 34(1), 28–40.

Vanderauwera, R. (1985) *Dutch Novels Translated into English: The Transformation of a „Minority“ Literature.* Amsterdam: Rodopi.

Védénina, L. G. (1980) La triple fonction de la ponctuation dans la phrase : syntaxique, communicative et sémantique. *Langue française* 45, 60–66.

Vinay, J.-P. and J. Darbelnet (1995) *Comparative Stylistics of French and English: A Methodology for Translation.* Amsterdam: John Benjamins.

Wachtarczyková, J. and R. Garabík (2016) Interlingválne faktory pri prechyľovaní cudzojazyčných ženských priezvisk v slovenčine. Časť 1. *Slovenská reč,* 81(3–4), 174–189.

CORPORA

Benko, V. *Araneum Anglicum Maius, version 15.04*. ÚČNK, Praha 2015. Available at: http://www.korpus.cz

Benko, V. *Araneum Bohemicum Maius, version 15.04*. ÚČNK, Praha 2015. Available at: http://www.korpus.cz

Benko, V. *Araneum Francogallicum Maius, version 15.03*. ÚČNK, Praha 2015. Available at: http://www.korpus.cz

The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. ÚČNK, Praha 2001. Available at: http://www.korpus.cz

Chlumská, L.: *JEROME: srovnatelný korpus překladové a nepřekladové češtiny*. ÚČNK, Praha 2013. Available at: http://www.korpus.cz

FRANTEXT corpus. ATILF: CNRTL. Available at: www.frantext.fr

Gaiffe, B. and K. Nehbi: *EstRepublicain, version 2*. ÚČNK, Praha 2016. Available at: http://www.korpus.cz

Klégr, A. et al.: *Korpus InterCorp — English, version 9 from 09/09/2016*. Ústav Českého národního korpusu FF UK, Praha 2016. Available at: http://www.korpus.cz

Křen, M. et al. *Korpus SYN, version 6 from 18/12/2017*. ÚČNK, Praha 2017. Available at: http://www.korpus.cz

Nádvorníková, O. and M. Vavřín: *Korpus InterCorp — French, version 10 from 01/12/2017*. ÚČNK, Praha 2017. Available at: http://www.korpus.cz

Rosen, A., M. Vavřín, and A. J. Zasina: *Korpus InterCorp — Czech, version 10 from 01/12/2017*. ÚČNK, Praha 2017. Available at: http://www.korpus.cz

Acknowledgements
This work was supported by the European Regional Development Fund-Project „Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World“ (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

This research was supported by the Charles University project Progres Q10, Language in the shiftings of time, space, and culture.

Olga Nádvorníková
Department of Romance Studies
Faculty of Arts, Charles University
Nám. Jana Palacha 2, 11638 Praha
ORCID ID: 0000-0001-7709-3901
olga.nadvornikova@ff.cuni.cz