sciendo
DOI 10.2478/jazcas-2025-0004

COMMA DISTRIBUTION IN CZECH TEXTS: VARIATION BY GENRE AND AUTHOR, AND ERROR ANALYSIS

JAKUB MACHURA[1] – HANA ŽIŽKOVÁ[2] – VOJTĚCH KOVÁŘ[3]
[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6623-3064)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6483-6603)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0009-0005-0307-9046)

MACHURA, Jakub – ŽIŽKOVÁ, Hana – KOVÁŘ, Vojtěch: Comma Distribution in Czech Texts: Variation by Genre and Author, and Error Analysis, 2025, Vol. 76, No 1, pp. 41 – 51.

Abstract: This article investigates the distribution and typology of commas in Czech texts, combining genre-differentiated samples with an annotated error corpus to offer a comprehensive view of punctuation usage and misuse. Building on previous work, we expand the analysis from a small newspaper sample to a broader set of texts, encompassing fiction, blogs, translations, and school dictations. Using a consistent typology of comma usage, we classify 1,000 manually selected instances and identify trends in different textual genres. Furthermore, we examine over 1,000 missing comma errors and more than 200 redundant ones from the self-built error corpus. The results reveal genre-dependent tendencies in comma types, especially in the use of commas preceding connectives and within asyndetic structures. The study offers insights for improving automatic comma insertion systems and deepens our understanding of punctuation norms and deviations in Czech.

Keywords: Comma typology, Punctuation errors, Czech language, Automatic comma insertion

1 INTRODUCTION

Punctuation plays a critical role in written communication, structuring sentences and guiding interpretation. Among punctuation marks, the comma is both frequent and frequently misused, making it a prime subject of computational linguistic research and grammatical annotation. In Czech, comma placement follows a complex set of syntactic rules and conventions, which are not always intuitively understood by writers—particularly in informal contexts or in translation.

A detailed typology of comma usage in Czech was proposed by Machura et al. (2022), laying the groundwork for further empirical analysis. The primary aim of the proposed typology was to systematically classify the positions of commas within Czech sentence structures, with particular attention to syntactic, semantic, and lexical factors. Such classification enables the identification of comma types that can be reliably defined through explicit linguistic rules, making them suitable for implementation in rule-based systems for automatic comma correction. Conversely, the typology also highlights those cases where comma placement is more ambiguous or context-dependent and thus requires statistical modeling or more advanced morphological, syntactic, and semantic analysis. While their initial study provided a valuable classification framework and a rough frequency estimate based on newspaper articles, the limited size and genre scope of the sample called for a broader, more representative dataset. At the same time, the need for improved evaluation methods for automatic comma insertion models has grown alongside the development of proofreader tools.

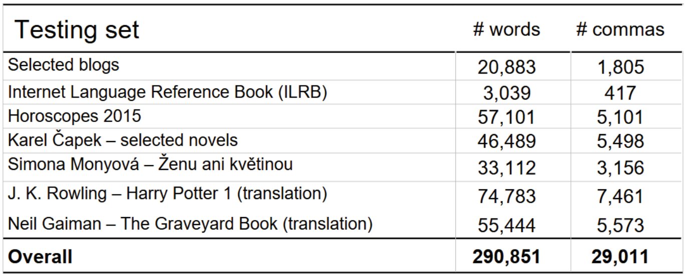
This article presents two complementary studies: first, a distributional analysis of 1,000 commas drawn from a variety of text genres; second, an error analysis using the self-built annotated corpus. Together, these perspectives illuminate both how commas are typically used in Czech writing and where writers most often go wrong.

2 COMMAS DISTRIBUTION IN CZECH TEXTS

In (Machura et al. 2022) the typology of comma insertion place was comprehensively described. This allows 1) to specify the place (boundary) in the sentence structure where a comma is inserted, 2) to analyze the type of commas that users of the language omit or overuse, or 3) to evaluate the results of language models that are pre-trained namely for the task of inserting commas into text, and then subsequently improve these models. Based on a relatively small sample of newspaper articles, which consisted of 183 sentence commas, a very rough frequency distribution of commas by type was outlined in that paper. Therefore, it was decided to analyze a larger sample that would more accurately determine the comma type distribution while also being representative, as it would consist of texts of different kinds, not just newspaper articles.

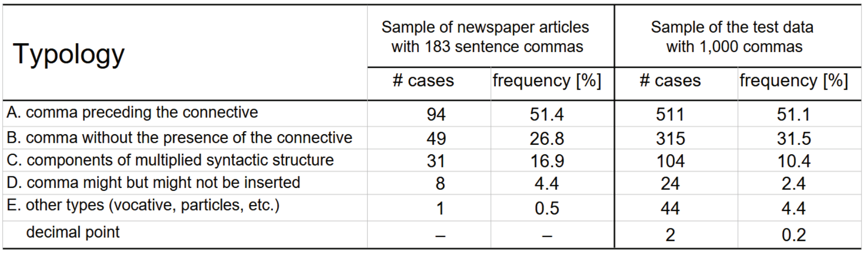
The new larger sample consisting of 1,000 commas was created from the same data presented in (Kovář et al. 2016), which are used specifically for the evaluation and comparison of methods for automatic comma insertion into Czech text. Since the data are exactly the same, it is also possible to compare the current results with testing done in the past (Machura et al. 2022; Machura et al. 2023). In total, seven texts of different natures and styles are used as testing data, see Tab. 1.

From each of the 7 texts, a sample containing 125 commas was selected. To add to the total of 1,000 commas, a sample from school dictations was included, which also contained 125 commas. All 1,000 commas were classified according to the selected typology and compared with a previous smaller sample (183 commas) from newspaper articles, see Tab. 2. The largest group, *A. a comma preceding the connective*, again reached slightly more than half of all commas (51.1%). Type *B. comma without the (near) presence of the connective* reached less than one-third (31.5%), while type *C. comma separating components of multiplied syntactic structure* decreased to only about one-tenth of all commas (10.4%). It turns out that type *D. cases where a comma is not obligatory or can change the meaning of the utterance* is even less frequent (2.4%), whereas type *E. commas around vocative phrases or particles and interjections* (standing outside the structure and syntactically independent) is more frequent (4.4%). This increase can also be explained by the selection of texts, as there were four fiction texts in the sample where vocative phrases may appear more often. There were also two commas in the sample which are used as a decimal point in numeral notation (in English, the symbol of the period is used as a decimal point whereas the comma also works as a thousand separator comma, and therefore both the period and the comma are ambiguous punctuation marks).



Testing set	# words	# commas
Selected blogs	20,883	1,805
Internet Language Reference Book (ILRB)	3,039	417
Horoscopes 2015	57,101	5,101
Karel Čapek – selected novels	46,489	5,498
Simona Monyová – Ženu ani květinou	33,112	3,156
J. K. Rowling – Harry Potter 1 (translation)	74,783	7,461
Neil Gaiman – The Graveyard Book (translation)	55,444	5,573
Overall	**290,851**	**29,011**

Tab. 1. Statistics of the test data for automatic comma insertion



Typology	Sample of newspaper articles with 183 sentence commas		Sample of the test data with 1,000 commas	
	# cases	frequency [%]	# cases	frequency [%]
A. comma preceding the connective	94	51.4	511	51.1
B. comma without the presence of the connective	49	26.8	315	31.5
C. components of multiplied syntactic structure	31	16.9	104	10.4
D. comma might but might not be inserted	8	4.4	24	2.4
E. other types (vocative, particles, etc.)	1	0.5	44	4.4
decimal point	–	–	2	0.2

Tab. 2. Comparison of the distribution of commas on a small (genre-specific) and a larger (genre-diverse) sample

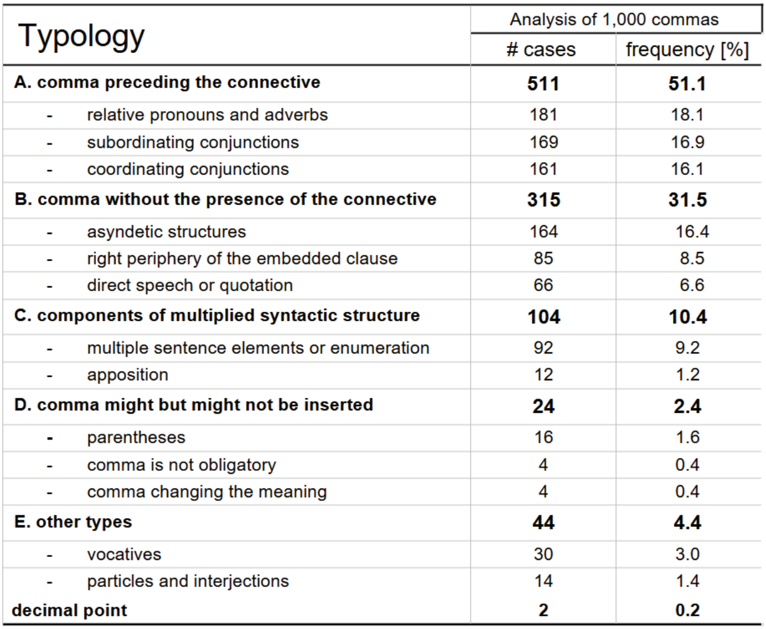
The table below presents a typological classification of 1,000 commas according to their syntactic function and context. Although this sample is not genre-balanced, the distribution confirms earlier findings about the predominance of type A commas, those preceding a connective, which account for 51.1% of all cases. Within this category, relative pronouns and adverbs (18.1%), subordinating conjunctions (16.9%), and coordinating conjunctions (16.1%) are represented in a relatively balanced manner, showing that various clause-linking strategies are equally comma-dependent in Czech syntax.

Type B commas, which appear without an explicit connective, form the second largest group (31.5%). Most notable within this type are asyndetic structures (16.4%), where elements are listed or juxtaposed without a linking word. Additionally, the right periphery of embedded clauses (8.5%) and direct speech or quotation (6.6%) reflect cases where comma placement relies more on syntactic and pragmatic cues than explicit connectives.

Type C commas, used in multiplied syntactic structures (e.g. enumerations and appositions), account for 10.4% of the sample. This relatively moderate proportion underscores the syntactic regularity of comma use in such constructions, with enumerations (9.2%) being more frequent than apposition (1.2%).

Optional commas (Type D) make up a small share (2.4%), typically found in parenthetical structures (1.6%), typically clauses with *"prosim" 'please'* or cases where punctuation can subtly alter the meaning or is simply not obligatory (0.4% each). This highlights the comparatively rare—but linguistically interesting—cases of stylistic or interpretative punctuation.

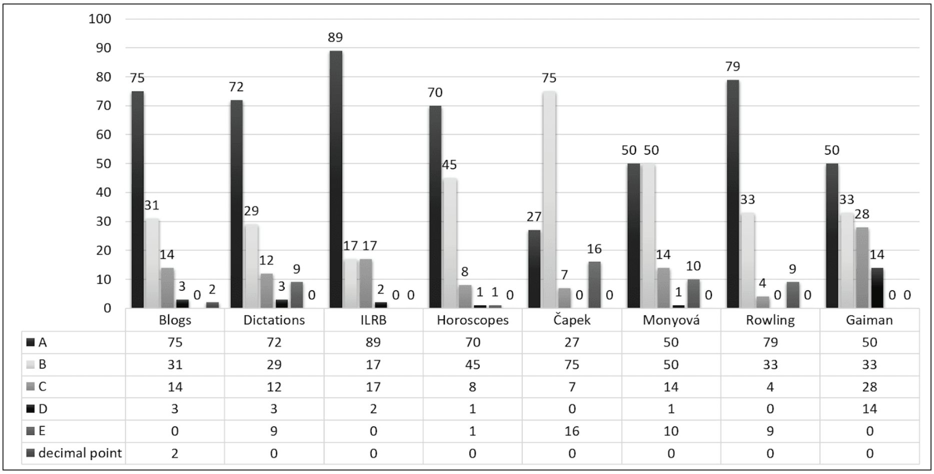
Other types (Type E) include vocatives (3.0%) and particles or interjections (1.4%), together forming 4.4%. These categories often fall outside the core syntactic structure and rely on discourse-level functions. Tab. 3 also includes decimal points (0.2%), which, while not true syntactic commas, are relevant for punctuation processing in computational contexts.



Typology	Analysis of 1,000 commas	
	# cases	frequency [%]
A. comma preceding the connective	**511**	**51.1**
- relative pronouns and adverbs	181	18.1
- subordinating conjunctions	169	16.9
- coordinating conjunctions	161	16.1
B. comma without the presence of the connective	**315**	**31.5**
- asyndetic structures	164	16.4
- right periphery of the embedded clause	85	8.5
- direct speech or quotation	66	6.6
C. components of multiplied syntactic structure	**104**	**10.4**
- multiple sentence elements or enumeration	92	9.2
- apposition	12	1.2
D. comma might but might not be inserted	**24**	**2.4**
- parentheses	16	1.6
- comma is not obligatory	4	0.4
- comma changing the meaning	4	0.4
E. other types	**44**	**4.4**
- vocatives	30	3.0
- particles and interjections	14	1.4
decimal point	**2**	**0.2**

Tab. 3. Observed distribution of 1,000 commas in detail

Different trends in comma distribution can be seen for each sample (see Tab. 4). Type *A. a comma preceding the connective* is prevalent for most texts, except for Čapek (21.6%, 27 commas) and Monyová (40%, 50 commas). Type B *comma* *without the (near) presence of the connective* is most frequent in Čapek (60%, 75 commas), Monyová (40%, 50 commas), in horoscopes this type is more common than in general (36%, 45 commas). However, sentences from the Internet Language Reference Book (2025) contain type B, which is far below average (13.6%, 17 commas). Type C is below average in horoscopes (6.4%, 8 commas), Čapek (5.6%, 7 commas) and Rowling (3.%, 4 commas). Gaiman, on the other hand, contains twice as many type C (22.4% with 28 commas) and more than 4 times as many type D (11.2%, 14 commas) as the average. Surprisingly, besides Gaiman, all samples of fiction texts contain type E, and the dictations contain an over-average of this type (7.2%, 9 commas; it can be assumed that this type was included in the dictations for didactic purposes).



	Blogs	Dictations	ILRB	Horoscopes	Čapek	Monyová	Rowling	Gaiman
■ A	75	72	89	70	27	50	79	50
■ B	31	29	17	45	75	50	33	33
■ C	14	12	17	8	7	14	4	28
■ D	3	3	2	1	0	1	0	14
■ E	0	9	0	1	16	10	9	0
■ decimal point	2	0	0	0	0	0	0	0

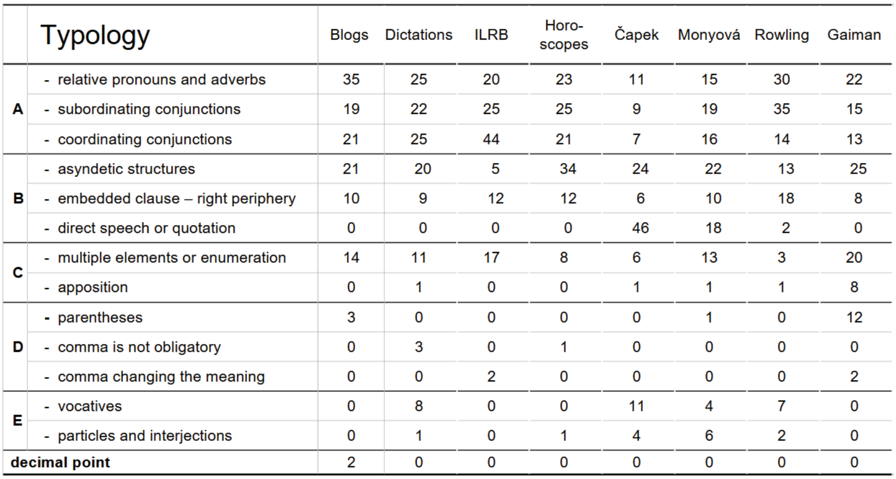
Tab. 4. Comparison of comma distribution in texts of different genres

Comparing the application of each sub-category across the samples provides a distinct perspective, see Tab. 5. The highest number of commas preceding relative clauses appears in blogs (35), while subordinating conjunctions are most frequently used in Rowling (35). Additionally, the ILRB example sentences contain the highest number of coordinating conjunctions (44). Hypotactic comma + connective is prevalent in all texts except ILRB, where the ratio of hypotactic to paratactic comma + connective is balanced (20 + 25 : 44). Čapek and Monyová, in particular, exhibit a significantly lower overall comma frequency throughout Section A.

The highest frequency of asyndetic structures, characterized by the absence of connectives, is found in horoscopes. Rowling's sample contains the greatest number of embedded sentences requiring separation at the right periphery (18). Additionally, more than two-thirds of all commas used around direct speech occur in Čapek (46). Notably, despite being a work of fiction, the Gaiman sample contains no instances of direct speech requiring the use of commas.

A more detailed analysis of type C. *components of multiplied syntactic structure* revealed that all samples contained instances of multiple elements or enumeration, while apposition appeared only marginally. Notably, it was entirely absent in blogs, ILRB, and horoscopes, whereas the Gaiman sample contained eight occurrences, which is relatively high given the small sample size. Similarly, nearly all instances of parentheses were found in Gaiman's text (12), with minimal representation in the other samples.

Commas marking vocatives were present in all fiction texts except for the Gaiman sample, with more than one-third occurring in Čapek's text (11). Additionally, eight instances of vocative commas were identified in dictation texts, where they likely were included deliberately for didactic purposes. The majority of commas surrounding particles and interjections also appeared predominantly in fiction, particularly in Monyová's text (6 instances).



	Typology	Blogs	Dictations	ILRB	Horo-scopes	Čapek	Monyová	Rowling	Gaiman
A	- relative pronouns and adverbs	35	25	20	23	11	15	30	22
	- subordinating conjunctions	19	22	25	25	9	19	35	15
	- coordinating conjunctions	21	25	44	21	7	16	14	13
B	- asyndetic structures	21	20	5	34	24	22	13	25
	- embedded clause – right periphery	10	9	12	12	6	10	18	8
	- direct speech or quotation	0	0	0	0	46	18	2	0
C	- multiple elements or enumeration	14	11	17	8	6	13	3	20
	- apposition	0	1	0	0	1	1	1	8
D	- parentheses	3	0	0	0	0	1	0	12
	- comma is not obligatory	0	3	0	1	0	0	0	0
	- comma changing the meaning	0	0	2	0	0	0	0	2
E	- vocatives	0	8	0	0	11	4	7	0
	- particles and interjections	0	1	0	1	4	6	2	0
decimal point		2	0	0	0	0	0	0	0

Tab. 5. Distribution of 1,000 commas in each subcategory

The sample sizes from individual authors are insufficient to provide conclusive insights into either the author or the overall syntactic structure of the text. However, they offer an indication of prevailing trends in individual texts. Furthermore, variations in the use of different types of commas can influence the effectiveness of automatic comma insertion, among other factors.

3 THE ERROR CORPUS

During the development of the online proofreader Opravidlo.cz, an error corpus was created. It is published in Sketch Engine (Killgariff et al. 2014) and can be searched by CQL queries. We used authentic texts from these domains: Autorevue.cz; Babinet.cz; Doktorka.cz; Hyperinzerce.cz; Seznamka.cz; Super.cz and Zpovednice.cz. Nine sets of hand-corrected sentences were compiled, each containing up to 2,000 sentences. Three annotators annotated each set, i.e. there are three versions of the corrections for each sentence. For a sentence to be included in the corpus, at least 2 of these 3 annotators had to agree on the correction, and the agreement had to be accurate to the letter (or 2 of the 3 had to say that the sentence was correct). With the agreement counted this way, we selected 13,829 sentences on which at least 2/3 agreed. This represents 82.5% (there were 2,939 disagreements). Of these, 4,136 sentences contain at least one tagged error, and the total set contains 6,411 tagged errors.

It should be added that annotators often marked errors in sentences by marking the wrong section in red. This method proved to be problematic in the subsequent evaluation of the sentences since the annotators used different editors (MS Excel, Libre Office, Google Docs). For this reason, these markings were ignored, and the error locations were calculated by comparing the error and the correction. For this reason, the section with an error is always the section between two spaces that contains the error, e.g. for punctuation errors, both the comma and the word before it are marked:

<s> Strašně mě << **vyděsilo | vyděsilo,** >> co se tu někde píše. </s>
'I was horrified by what is written somewhere here.'

On the one hand, this is a rather primitive practice, and it might be worth marking it more precisely, at least in some cases; on the other hand, it is somewhat consistent with the idea that the proofreader tool being developed should instead underline some larger section of text so that the warning is visible.

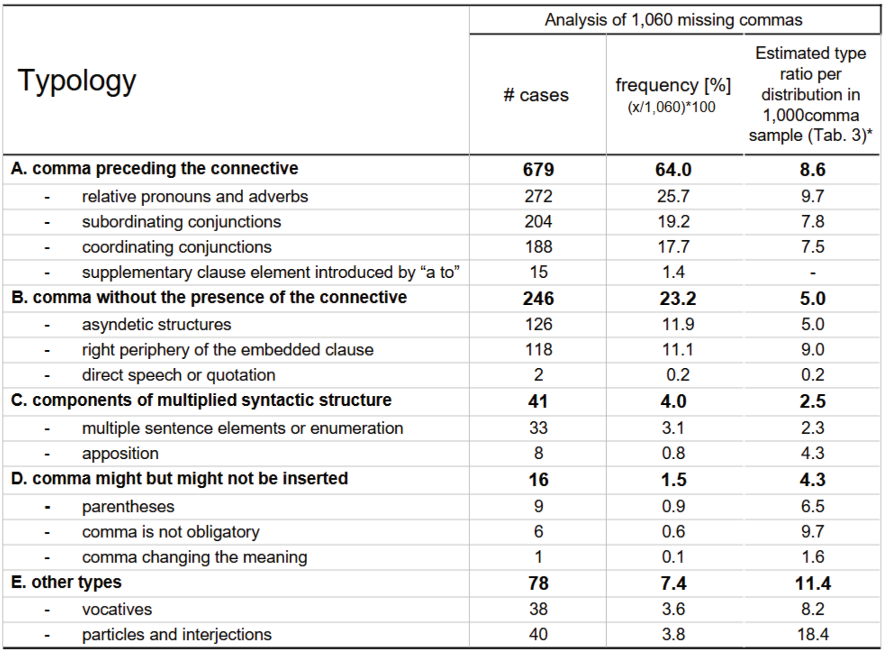
When analyzing the annotated sentences, some systematic problems became apparent, for example, roughly one-tenth of all marked errors are corrections of hyphenation to hyphen, which may be due to the normalization of the texts. The situation is similar to other typographical errors such as quotation marks or other characters.

3.1 Punctuation errors

The *Error corpus* contains a total of **15,516 commas**, including 19 decimal points, leaving **15,497 valid sentence commas**. In total, there are 1,066 corrections where a comma was added after a word (found using CQL: `(<err/> !containing ",")(<corr/> containing ","))`. Of these, **1,060 instances** were analyzed as **missing comma errors**. This means that the writers of these blog texts achieved a **recall (R) of 93.2%** for correctly written sentence commas (R = 14,437 / (14,437 + 1,060)).

Conversely, using the query `(<err/> containing ",")(<corr/> !containing ",")`, there are 247 corrections where a comma was removed. In **218** of these cases, the comma was identified as **redundant**. If we assume that the writers produced 14,437 commas correctly and 218 commas redundantly, their **precision (P)** would be **98.5%** (P = 14,437 / (14,437 + 218)). The relatively high recall and precision can be related to the simpler sentence structure of the blogs where type *A. a comma preceding connective* usually prevails and writers do not usually have problems writing a comma before a connective.

In a more detailed analysis of the missing commas (see the following Tab. 6), two-fifths of the missing commas of type A represent a comma before relative pronouns and adverbs (25.7% of all missing commas, 272 commas), which in Czech usually separate relative clauses from the rest of the sentence. In 204 cases (19.2% of all missing commas), writers omitted a comma before subordinating conjunctions that separate the subordinate clause from the main clause. To a slightly lesser extent, the writers failed to insert a comma before the connective that separates sentences that are formally coordinated (17.7%, 188 commas). If we consider the distribution of commas according to Tab. 3, then writers had slightly more difficulty placing commas before relative pronouns and adverbs (they omitted 9.7% of commas that should be placed before relative pronouns and adverbs). The last group, which cannot be fully classified under the previous three, consists of supplementary clause elements introduced by *"a to" 'and this'* (1.4%, 15 commas).



Typology	Analysis of 1,060 missing commas		
	# cases	frequency [%] (x/1,060)*100	Estimated type ratio per distribution in 1,000comma sample (Tab. 3)*
A. comma preceding the connective	**679**	**64.0**	**8.6**
- relative pronouns and adverbs	272	25.7	9.7
- subordinating conjunctions	204	19.2	7.8
- coordinating conjunctions	188	17.7	7.5
- supplementary clause element introduced by "a to"	15	1.4	-
B. comma without the presence of the connective	**246**	**23.2**	**5.0**
- asyndetic structures	126	11.9	5.0
- right periphery of the embedded clause	118	11.1	9.0
- direct speech or quotation	2	0.2	0.2
C. components of multiplied syntactic structure	**41**	**4.0**	**2.5**
- multiple sentence elements or enumeration	33	3.1	2.3
- apposition	8	0.8	4.3
D. comma might but might not be inserted	**16**	**1.5**	**4.3**
- parentheses	9	0.9	6.5
- comma is not obligatory	6	0.6	9.7
- comma changing the meaning	1	0.1	1.6
E. other types	**78**	**7.4**	**11.4**
- vocatives	38	3.6	8.2
- particles and interjections	40	3.8	18.4

Tab. 6. Observed distribution of the missing 1,060 commas in the Error corpus

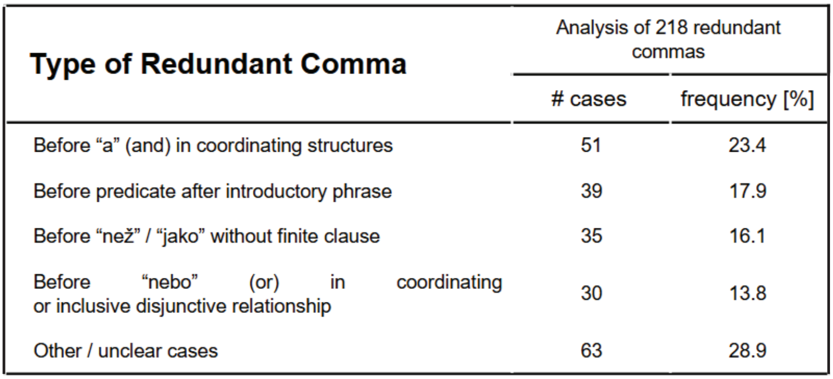
*E.g., the Error corpus is expected to contain approximately 15,500 correctly placed sentence commas. Of these, an estimated 51.1% are of type A (see Tab. 3), which corresponds to about

7,920 commas. Out of this subset, 679 commas — or 8.6% of type A — were omitted by the writers.

More than half of the missing commas of type B are asyndetic structures with no presence of a connective (11.9%, 126 commas). Embedded clauses usually contain a connective on the left side, but are separated asyndetically from the right. In these cases, the comma was missing 118 times (11.1%). This type appears to be more problematic, as we estimate that the writers did not close 9% of the embedded clauses from the right periphery.

The lowest number of errors was recorded for types C (4.0%, 41 commas) and D (1.5%, 16 commas). Almost uniformly, commas were missing around vocative phrases (3.6%, 38 commas) and particles and interjections (3.8%, 40 commas). Writing commas around particles and interjections appears to be the most difficult for writers (they forgot to insert a comma in 18.4% of cases), whereas they only forgot commas around vocatives in 8.2% of cases.

A closer look at the 218 cases of redundant commas (see Tab. 7) reveals several recurring patterns. The most frequent error type (23.4%, 51 instances) was the insertion of a **comma before the conjunction *a* (‘and’) in coordinating structures**, where no comma is required. Surprisingly, the second most common type (17.9%, 39 instances) involved a **comma erroneously placed between the initial phrase and the predicate** — e.g. *“Dům s praktickou dispozicí, nabízí příjemné bydlení”* ‘*A house with a handy layout provides a comfortable living experience’* or *“Z hlediska praktičnosti využití jeho patentů*, má ohromný náskok před Edisonem”* ‘*In terms of the practical use of his patents, he has a significant advantage over Edison’*. These commas may reflect either a prosodic pause (as in spoken language) or influence from English syntactic patterns (e.g. introductory adverbs or phrases).



Type of Redundant Comma	Analysis of 218 redundant commas	
	# cases	frequency [%]
Before “a” (and) in coordinating structures	51	23.4
Before predicate after introductory phrase	39	17.9
Before “než” / “jako” without finite clause	35	16.1
Before “nebo” (or) in coordinating or inclusive disjunctive relationship	30	13.8
Other / unclear cases	63	28.9

Tab. 7. The most common types of redundant commas

Another common issue, observed in 35 instances (16.1%), was a **redundant comma before the conjunctions *než ‘than’* and *jako ‘as’***. In Czech, these conjunctions only require a comma if they are followed by a finite verb clause. Omitting this distinction often leads to unnecessary punctuation. The last more frequent group (13.8%, 30 commas) was the **redundant comma before the conjunction *nebo* ‘*or*’ in a coordinating or inclusive disjunctive relationship** (in Czech, the comma before *nebo* is written when using any of correlative conjunctions such as *ať–nebo* ‘whether–or’*, buď–nebo* ‘either–or’ or in exclusive disjunction). The remaining redundant comma cases were less frequent and often lacked a clear syntactic or prosodic motivation.

4 CONCLUSION

This study offers a comprehensive analysis of comma usage in Czech texts, integrating typological classification with distributional and error analyses. The expanded sample of 1,000 classified commas corroborates previous findings regarding the predominance of commas preceding connectives while also emphasizing genre-specific variation—particularly in fiction, where commas not accompanied by a connective, as well as those marking vocatives and syntactically independent expressions, occur more frequently. In the next phase of research, it would be useful to compare comma distribution with other genres (primarily non-fiction).

The analysis of the Error corpus further reveals systematic patterns in punctuation errors. Writers most commonly omit commas before relative pronouns and subordinating conjunctions or within asyndetic structures, while redundant commas often appear in positions influenced by prosody or interference from English syntax. Despite these challenges, the high overall precision and recall of comma usage in informal web texts suggests a strong intuitive grasp of fundamental rules among Czech writers. These findings not only enhance our understanding of punctuation norms in Czech but also provide valuable feedback for the development of automated comma insertion tools.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was supported by the OSCARS project, funded by the European Commission’s Horizon Europe Research and Innovation programme (grant agreement No. 101129751), led by the five Science Clusters: ENVRI, ESCAPE, LS RI, PaNOSC, and SSHOC.

References

Hlaváčková D. et al. (2022). Opravidlo.

Internet Language Reference Book. (2025). Praha: Ústav pro jazyk český AV ČR.

Kilgariff, A. et al. (2014). The Sketch Engine: ten years on. Lexicography. Springer Berlin Heidelberg, 1(1) pp. 7–36. Accessible at: https://dx.doi.org/10.1007/s40607-014-0009-9.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts: Comparative Study. Online. In: P. Sojka – A. Horák – I. Kopeček – K. Pala (eds).: Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic (September 6 – 9, 2022) Proceedings. Cham (CH): Springer, pp. 113–124. Accessible at: https://dx.doi.org/10.1007/978-3-031-16270-1_10.

Machura, J. et al. (2023). Is it Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction. Jazykovedný časopis, (74)1, pp. 357–368. Accessible at: https://dx.doi.org/10.2478/jazcas-2023-0052.