sciendo
DOI 10.2478/jazcas-2023-0052

IS IT POSSIBLE TO RE-EDUCATE ROBERTA? EXPERT-DRIVEN MACHINE LEARNING FOR PUNCTUATION CORRECTION

JAKUB MACHURA[1] – HANA ŽIŽKOVÁ[1]
– ADAM FRÉMUND[2] – JAN ŠVEC[2]
[1] Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

MACHURA, Jakub – ŽIŽKOVÁ, Hana – FRÉMUND, Adam – ŠVEC, Jan: Is it Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 357 – 368.

Abstract: Although Czech rule-based tools for automatic punctuation insertion rely on extensive grammar and achieve respectable precision, the pre-trained Transformers outperform rule-based systems in precision and recall (Machura et al. 2022). The Czech pre-trained RoBERTa model achieves excellent results, yet a certain level of phenomena is ignored, and the model partially makes errors. This paper aims to investigate whether it is possible to retrain the RoBERTa language model to increase the number of sentence commas the model correctly detects. We have chosen a very specific and narrow type of sentence comma, namely the sentence comma delimiting vocative phrases, which is clearly defined in the grammar and is very often omitted by writers. The chosen approaches were further tested and evaluated on different types of texts.

Keywords: comma, Czech, vocative, machine learning, RoBERTa

1 INTRODUCTION

Punctuation, along with other graphical features, is a key element for the correct understanding of any text. Nunberg (1990) describes punctuation as a linguistic subsystem of written language that identifies different text categories. Among the punctuation marks, the sentence comma generally receives the biggest attention. Its use is most often described in grammar, because it is the most frequent punctuation mark across various languages (see Švec et al. 2021; Chordia 2021).

Two tasks usually deal with the automatic insertion of a sentence comma: (1) punctuation restoration in speech transcripts from automatic speech recognition (ASR), where the restored punctuation dramatically improves the readability of the recognised transcript, and (2) grammatical error correction in written texts where the comma is missing or redundant.

For the first task, defining the beginning and the end of the sentence unit is necessary. A system should handle capitalisation to identify the beginning and insert a period or question mark to indicate the end of the sentence. The comma itself then serves as the intra-sentence mark that helps to structure the sentence on syntactic and semantic levels. The second task usually does not have to deal with the insertion of terminal punctuation marks, as these usually already appear in the texts. It is obvious both tasks have different starting positions, and we do not consider it fair to compare the results of tools that insert commas in texts for different tasks.

Nevertheless, the BERT-based method for ASR proposed in (Švec et al. 2021) had promising results, so we decided to train a new model targeted only on the comma correction task (hereinafter referred to as **RoBERTa-commas**). The new model's results were presented in (Machura et al. 2022). The precision was slightly higher (96.1%) than the best rule-based system for Czech (95.1%). But the new model completely outperforms the recall (number of correctly found commas) – 89.5% versus 58.8%. Although the results of the new model are excellent, for the purposes of a reliable language corrector, we were interested to find out whether we would be able to modify the model in any way further – to find and correct the errors it makes and to identify the 10% of commas that the model ignored. The goal of this paper is to see whether we can expand the number of commas found by modifying the learning process. Two new models – **Extra-Fine-tuning** and **Mixed-data-fine-tuning** – will be compared with the **RoBERTa-commas** model.

2 APPROACH

Writing commas in Czech is traditionally part of the grammar, and reference books extensively describe the rules for writing commas. Nunberg (1990) recognises two main classes of commas. He sees a difference between a comma which separates structures at the same level (the separator comma), and a comma which delimits the boundaries between syntactic structures at different levels (the delimiter comma). To define all possible places where a comma should be inserted in the Czech text, we have created a typology of the comma insertion places described in detail (Machura et al. 2022).

In our previous research, we used a corpus of newspaper articles with 87,379 commas for further analysis. All commas were intentionally deleted, and the **RoBERTa-commas** model reinserted commas back into the texts. From the texts, we selected a sample of about 10,000 sentences that did not contain a comma, or a comma was inserted in a different place compared to original newspaper articles. These sentences are currently being further classified according to the established typology. As a result, we can more clearly identify types of commas that are ignored or incorrectly inserted by the model.

When we find a comma that the model did not insert into the text, the first question is why. Is it a type that was not included in the training data, and therefore the model could not learn it? And if so, can we influence the training process in any way – e.g. by making the training data larger or by extra fine-tuning? Or is there a possibility to do postprocessing and fix the commas using formal linguistic rules?

Based on these questions, we decided to choose a very narrow class of commas, namely commas around vocatives ("vocative comma"), for the following reasons:

1. it is a very common error among a large number of writers[1]; 2. writing of commas around vocative in Czech has very fixed rules without any exceptions (Internet Language Reference Book 2023).

We wondered whether it was possible to improve the language model by supplying it with specific sentences containing constructions that were previously evaluated incorrectly by the model.

2.1 Vocative

Traditionally, the vocative is understood as a case that significantly differs from others. Morphologically, it behaves like other cases in Czech – the vocative suffix is attached to the right periphery of the noun. However, the vocative is not governed by any other clause element. In the vast majority of cases, it is a noun phrase that is not part of the valence frame. Its position in the clause is almost arbitrary – at the beginning: *Pane, vyslyš nás.* 'Lord, hear us.'; in the middle: *Prosím, pane, vyslyš nás.* 'Please, Lord, hear us.'; and at the end: *Vyslyš nás, pane.* 'Hear us, Lord.' Some theories propose the idea that the vocative stands completely outside the syntactic structure of the clause – the evidence is, e.g. the vocative's invisibility to clitics that compulsorily occupy the second position: ***Karle, se schovej!* × *Karle, schovej se!*** 'Charles, hide!' (Karlík 2017).

Some suffixes of masculine singular (e.g. *Bože* 'God' etc.) and feminine singular (e.g. *ženo* 'wife') are unique only to the vocative. However, most vocative suffix forms are shared with other cases, most often with the nominative case (e.g. *paní ministryně* 'Madame Minister' etc.). Therefore, the presence of a comma around the vocative is crucial for automatic morphological analysis if the vocative is to be correctly tagged. This is also one of the reasons why rule-based systems dependent on morphological analysis cannot detect vocative commas reliably.

[1] For instance, a probe of the web corpus Araneum Bohemicum Maius (Benko 2015) confirmed the error rate of the vocative forms for the noun *slečna* 'Miss' and the proper noun *Adam*. The corpus contains 635 occurrences of the vocative *Slečno* at the beginning of a sentence of which 33 instances (5%) are missing a comma (e.g. *Slečno nemáte tam něco ostřejšího?* 'Miss do you have anything sharper?'). The form *Adame* is found 393 times in the corpus, with at least 59 occurrences (approx. 15%) being vocatives without a missing comma.

3 ROBERTA-COMMAS TRAINING

We used our pre-trained model on a collection of web data processed in our web mining tool (Švec et al. 2014), CommonCrawl data[2] and texts collected from the Czech Wikipedia. We followed all the training steps mentioned in (Liu et al. 2019) for pre-training the Czech RoBERTa model (Lehečka et al. 2021). As suggested, we used dynamic masking of the tokens. This strategy generates the masking pattern every time a sequence is fed to the model. Also, the pre-training procedure does not use Next Sentence Prediction loss in contrast with the BERT model (Devlin 2019). For tokenizing the input sequence, we used Byte-Pair Encoding (BPE), introduced in (Radford et al. 2019), with a subword vocabulary containing 50K units. This implementation of BPE uses bytes instead of Unicode characters as the base subword units.

We used the ADAM optimisation with linear warmup up to learning rate $4 \cdot 10^{-4}$ for the first 24K steps followed by linear decay to 0. We pre-trained the model for 500K steps as described in (Lehečka 2021).

For this experiment, we proposed the RoBERTa model extended by a few extra classification layers (Švec et al. 2021). An input sequence is preprocessed using the tokenizer, and special tokens are added. Next, we use the Czech pre-trained RoBERTa model with output dimension d=768. The last hidden states are transformed by four regular densely-connected neural network layers. Three of these layers use the element-wise ReLU activation function, and the last layer uses the softmax activation function. The last layer output defines whether the comma should be placed right after the current token. The overall scheme of the proposed neural network architecture is depicted in Fig. 1.

As the training data set used for fine-tuning, we used 10 GB of raw text extracted from the Czech CommonCrawl data set. Because RoBERTa's output is related to input tokens (not words), we assigned the target label ("," for comma, "0" for no comma) to each token of the word, which is followed by a comma (as shown at Fig. 1). In the prediction phase, it is necessary to combine the predictions for the partial tokens into word-level predictions using a per-word pooling. We use a simple average pooling algorithm to obtain the word-level predictions. As the model defines this experiment as a two classes classification per each token, we use standard categorical cross-entropy loss. For optimisation, we use the ADAM optimisation algorithm. The parameters of the whole network (**RoBERTa-commas**) – consisting of the RoBERTa and the classification layers – were updated during fine-tuning.

The epoch size of **RoBERTa-commas** equals 10K sequences, the batch size equals 45, and the number of epochs equals 25, 50 and 75. During the fine-tuning, we use a linear learning rate decay with values starting at 10^{-4} and ending at 10^{-5}. The maximum sequence length is set to 128.

[2] https://commoncrawl.org/

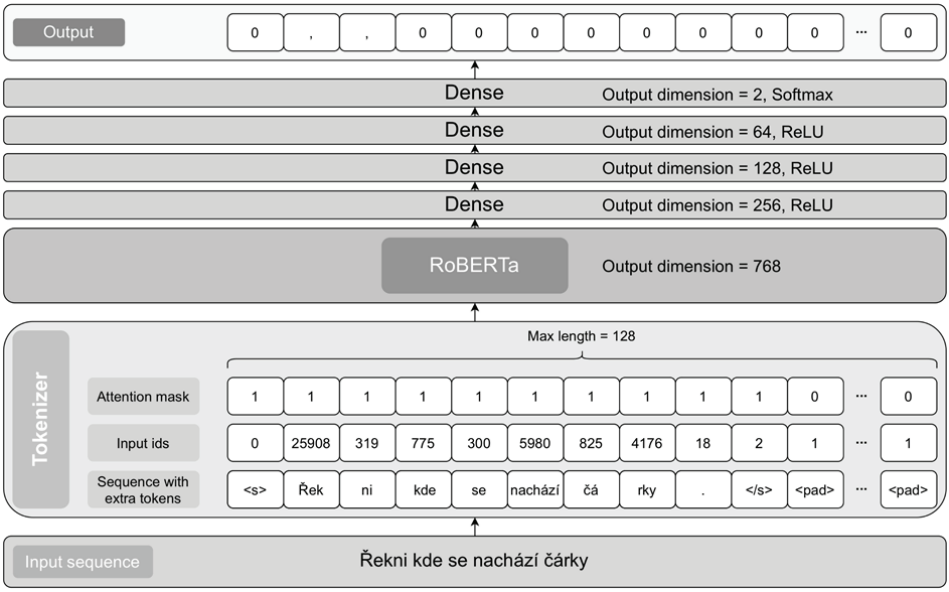


Fig. 1. Proposed neural network architecture

Then, the analysis of the incorrectly evaluated sentences followed (on approximately 10K sentences), and a sample of 89 sentences that did not contain a comma around the vocative was selected. The language model is quite successful regarding proper nouns and nouns. Nevertheless, the model did not insert some commas, mostly for appellative vocatives:

Kluci já jsem delegace* 'Guys* I am a delegation'.

To chce klid vládo, že ano?* 'Take it easy* government, won't you?'

There were also instances where the model inserted a comma in the wrong place.

> *Paní ministryně*, kariérně prošla Senát.* 'Mrs. Minister*, has had a career in the Senate.'

We assume the model made this error because the form *paní ministryně* 'Mrs. Minister' appeared very often as a vocative in the training data. Therefore, we were interested in identifying the most frequent appellative Czech vocative.

To further improve the **RoBERTa-commas** model after examining the phrases that had been improperly processed, we decided to build a small extended text corpus that purposely contained vocative phrases. The following section goes into further detail about this corpus.

Firstly, we attempted to improve the **RoBERTa-commas** model that had already been fine-tuned using the corpus of vocative phrases. Thus, we used a model that was fine-tuned on 10 GB of raw text data, and then we ran a quick extending

(second) fine-tuning process (utilizing the corpus of vocative phrases) to adjust the weights of the RoBERTa-commas model. This approach will hereafter be referred to as **Extra-Fine-tuning**.

Second, we tried a different approach by merging the original raw text corpus with the corpus of vocative phrases. Then, using the newly obtained mixed data, we replaced the initial fine-tuning procedure on raw text data with this mixed data. This approach will be referred to as the **Mixed-data-fine-tuning**. Both approaches are described in more detail in the following sections.

3.1 Retraining data collection

We performed a corpus analysis on the csTenTen17 corpus (Suchomel 2018), which is a web corpus containing 10.5 billion words. It turned out that the appellative vocative in Czech consists most often of one to three words (in the case of four or more consecutive vocative word forms, these were enumerations, repeated exclamations, etc.). In the case of one-word vocative, the most frequent lemmas are *pan, bůh, doktor* 'mister, god, doctor'. If it is a two-word vocative, constructions such as *pane ministře, pane doktore, vážený pane* 'Mr. Minister, Mr. Doctor, Dear Sir' are most frequent. Feminine pronouns are less frequent. Furthermore, it turns out that appellative two-word vocatives often have a form with a personal pronoun: *ty vole, ty bláho, vy idiote, vy troubo* 'you dude, you fool, you idiot, you moron'. In the case of vocative containing three parts, the most frequent are *vážený pane přísedící, vážený pane místopředsedo, vážený pane předsedo* 'dear Mr. Chair, dear Mr. Vice-President, dear Mr. Chairman'.

Based on the analysis, we collected a set of 170,000 sentences with 82,381 vocatives using CQL queries.[3] The set contained sentences where there were one to three words in the vocative between punctuation marks. Proper nouns were filtered and removed. Since the previous set completed by the language model lacked punctuation for feminine vocatives (the type *paní ministryně* 'Mrs. Minister'), we tracked down sentences that explicitly contained the lemma *paní* 'madame'. A sample of selected sentences was as follows:
one-word vocative – 30,000 sentences
two-word vocative – 30,000 sentences (including approximately 18,000 sentences with vocative type *pane ministře* 'Mr. Minister', and 12,000 sentences with vocative type *paní ministryně* 'Mrs. Minister'
three-word vocative – 11,271 sentences
type *vy osle* 'you dude' pl. – 1,110 sentences
type *ty osle* 'you dude' sg. – 10,000 sentences

In each sentence, the vocative was separated by punctuation, and the position of the vocative in the structure varied.

[3] Such as [tag="kI.*"][tag=".*c5.*"][tag="kI.*"] and its modifications.

4 EXPERIMENTAL RESULTS

4.1 Findings

Both approaches (Extra-Fine-tuning and Mixed-data-fine-tuning) for learning the "vocative comma" type were first evaluated on a corpus of newspaper articles. As mentioned above, the sample of 89 sentences containing mainly appellative vocatives was used to analyse this specific type. **Extra-Fine-tuning** brought significantly better results. More than 60% of the vocative phrases in the sample were correctly delimited, whereas the **Mixed-data-fine-tuning** correctly delimited less than 40%.

Vocative – total 89	**Extra-Fine-tuning**	**MIXED-DATA-FINE-TUNING**
Correctly delimited	54	35
Ignored	22	38
Delimited only from one side	5	7
Semantically ambiguous	8	9

Tab. 1. Evaluation of both approaches (Extra-Fine-tuning and Mixed-data-fine-tuning) on a sample of 89 vocatives

The sample also contained semantically ambiguous clauses, and there was more than one option for inserting a comma.

E.g.
Ale jen generační, to není jasný, pane. 'But only generational, that's not clear, sir.'
vs.
Ale jen generační to není, jasný pane. 'But it's not just generational, kind sir.'

The data used to retrain the model also affected other types of sentence commas. The RoBERTa-commas model could not handle tokens that were connected with other punctuation marks in the text (the original training data contained only periods, commas and question marks):

Tedy představte si sloupec 70 cm vody (po ***stehna) kdyby*** *to spadlo naráz.* 'So imagine a column of 70 cm of water (up to your thighs) if it fell all at once.'

Ilf zmiňuje ve svých zápiscích ***„tvář jež*** *nejeví vyčerpanost z mentálního úsilí".* 'Ilf mentions in his notes "a face that does not show exhaustion from mental effort".'

However, the data used for retraining also contained other punctuation marks. As a result, during the **Extra-Fine-tuning**, the model learned to insert commas after tokens that were directly connected with another punctuation mark:

*Tedy představte si sloupec 70 cm vody (po **stehna), kdyby** to spadlo naráz.*

*Ilf zmiňuje ve svých zápiscích „**tvář, jež** nejeví vyčerpanost z mentálního úsilí“.*

And the **Mixed-data-fine-tuning** model no longer incorrectly inserts a comma in a sentence with the homonymous phrase *paní ministryně* ‘Mrs. Minister’ mentioned above:

Paní ministryně kariérně prošla Senát. ‘Mrs. Minister has had a career in the Senate.’

Apparently, retraining can improve the detection of a specific type if we intentionally control what the model has to learn. This small sample of vocatives, though, is a part of the corpus of newspaper articles that has 87,379 commas, and in general, we estimate that vocative is a very marginal phenomenon, less than 1% of all commas of text (Machura et al. 2022). Therefore, we were interested in how retraining one specific type affects the global metrics on the entire corpus, which is used for detailed analysis of the insertion of the automatic comma.

Evaluation of precision, recall and F1 score (the harmonic mean of precision and recall) for the **RoBERTa-commas**, **Extra-Fine-tuning** and **Mixed-data-fine-tuning** on the corpus of newspaper articles is presented in the following table.

Evaluation on the corpus of newspaper articles total number of commas: 87,379					
	Absolute number of commas inserted	**Number of commas correctly inserted**	**P [%]**	**R [%]**	**F1 [%]**
RoBERTa-commas	78,146	76,250	97.6	87.3	92.1
Extra-Fine-tuning	77,976	75,839	97.3	86.8	91.7
Mixed-data-fine-tuning	77,890	75,553	97.0	86.5	91.4

Tab. 2. Evaluation of precision, recall and F1 score

While the evaluation showed that both approaches improved “vocative comma” detection, it seems that the retraining data slightly degraded all global metrics. This may be caused partly by the fact that “vocative comma” is a very marginal phenomenon, and the retraining data was too large (more than 80K sentences) and distorted the natural distribution of all types of commas in the original training data. With this hypothesis in mind, we decided to evaluate both approaches on real texts of different natures.

4.2 Evaluation data sets

The same data as presented in (Kovář et al. 2016) were used to evaluate and compare the methods described above. These texts were prepared specifically for the automatic insertion of commas. Since the data are exactly the same, it is also possible to compare the current results with testing done in the past. In total, seven texts of different nature and styles were used, see Tab. 3.

Testing set	**# words**	**# commas**
Selected blogs	20,883	1,805
Internet Language Reference Book	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. 3. Statistics of the test data

Both newly trained models behaved differently than we expected (see Tab. 4). They significantly improved precision, with **Mixed-data-fine-tuning** outperforming **Extra-Fine-tuning**. This fact disproved our original hypothesis, as we predicted that precision would decrease based on the results from the corpus of newspaper articles (see Tab. 2).

Recall, on the other hand, decreased more for both models than we expected. For the **Extra-Fine-tuning** by 4.5 percentage points, for the **Mixed-data-fine-tuning** even by 11.5 percentage points. For lower-register texts (blogs and horoscopes), the **Mixed-data-fine-tuning** found more correct commas than the **Extra-Fine-tuning**, whereas, for fiction, the **Extra-Fine-tuning** won over the **Mixed-data-fine-tuning**.

A possible explanation might be that stylistically lower texts would not be rich in complex constructions, and writers use frequent connectors and basic sentence structures. On the other hand, fiction is full of direct speech and more complex sentence constructions, and this requires more frequent delimitation of parts of the text without the presence of connectors. The original model outperformed both retrained models in recall and F1 score. The absolute number of inserted commas clearly shows that RoBERTa-commas tried to insert far more commas than the two new models (Mixed-data-fine-tuning even inserted almost 4,000 fewer commas than RoBERTa-commas). Therefore, a logical explanation for the better precision is

offered: The fewer commas a model adds, the less chance it has to make a mistake. We deliberately did not compare our trained models with other approaches that deal with automatic insertion of commas because this would not answer our research question.

RESULTS	***ROBERTA-COMMAS***			***EXTRA-FINE-TUNING***			***MIXED-DATA-FINE-TUNING***		
	P [%]	**R [%]**	**F1 [%]**	**P [%]**	**R [%]**	**F1 [%]**	**P [%]**	**R [%]**	**F1 [%]**
SELECTED BLOGS	**95.5**	**88.3**	**91.8**	**97.5**	**83.3**	**89.8**	**97.6**	**84.8**	**90.8**
INTERNET LANGUAGE REFERENCE BOOK	**91.8**	**70.0**	**79.5**	**95.5**	**55.4**	**70.1**	**96.1**	**52.5**	**67.9**
HOROSCOPES 2015	**96.4**	**93.9**	**95.2**	**97.5**	**87.7**	**92.3**	**97.7**	**87.7**	**92.4**
KAREL ČAPEK – SELECTED NOVELS	**95.3**	**88.9**	**92.0**	**97.3**	**85.7**	**91.1**	**97.5**	**78.8**	**87.2**
SIMONA MONYOVÁ – ŽENU ANI KVĚTINOU	**95.8**	**93.1**	**94.4**	**97.5**	**88.6**	**92.8**	**98.3**	**86.3**	**91.9**
J.K. ROWLING – HARRY POTTER 1 (TRANSLATION)	**96.6**	**88.4**	**92.3**	**97.6**	**85.0**	**90.9**	**97.9**	**77.5**	**86.5**
NEIL GAIMAN – THE GRAVEYARD BOOK (TRANSLATION)	**96.8**	**87.2**	**91.8**	**98.0**	**82.1**	**89.4**	**98.6**	**64.3**	**77.8**
OVERALL PERFORMANCE	**96.1**	**89.5**	**92.7**	**97.6**	**84.9**	**90.8**	**97.9**	**78.0**	**86.9**
Absolute number of commas inserted	**27,000**			**25,244**			**23,125**		
Number of commas correctly inserted	**25,958**			**24,631**			**22,640**		

Tab. 4. Results: RoBERTa-commas, Extra-Fine-tuning and Mixed-data-fine-tuning

5 CONCLUSION

This paper aimed to find out whether it is possible to re-train the language model RoBERTa by providing example sentences in which the language model made errors. We performed a very narrowly focused study: we chose to detect commas in the Czech vocative. We deliberately chose an area that is not problematic in Czech, and writing commas, in this case, is fixed without exceptions. We analysed the types of Czech sentences with vocative, determined how such sentences look in an authentic text and developed a typology, according to which we collected 170,000 sentences with 82,381 vocatives from the csTenTen17 corpus using CQL queries. We used this set of sentences to retrain RoBERTa-commas.

If we precisely analyse the phenomenon that the RoBERTa-commas model ignores by means of retraining, we can detect a specific type. For retraining, we created a corpus that purposefully contained commas around appellative vocatives. We then tried to integrate this new corpus into the learning process: (1) as extra additional training (**Extra-Fine-tuning**), (2) we merged the specialised corpus with the original large corpus of training data (**Mixed-data-fine-tuning**). However, testing showed that in our case, deliberate modification of the original training data or extra fine-tuning significantly reduced recall and likely led to an increase in precision.

This result is probably caused by overfitting the neural network to one specific type of comma. The model weights are then adjusted to best cover the comma around the vocative, and all other types of commas could be overlooked. As the neural network is more like a black box, we cannot say with certainty that this is the case. But it could be a good idea to consider the distribution of the phenomena in the text and proportionally create the retraining data accordingly. The comma types distribution has already been roughly estimated in (Machura et al. 2022). Still, this estimate was made on a very small sample, and the distribution will need to be calculated on a larger sample of different types of texts. Otherwise, the new knowledge may overshadow the original functionality.

Our analysis showed that re-training RoBERTa is possible, but the structure of training data plays an important role.

ACKNOWLEDGEMENTS

A. Frémund and J. Švec were supported by the grant of Czech Science Foundation (GA CR), project No. GA22-27800S.

References

Benko, V. (2015). Araneum Bohemicum Maius, verze 15.04. Ústav Českého národního korpusu FF UK, Praha 2015. Accessible at: http://www.korpus.cz.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. /abs/1810.04805 Accessible at: https://doi.org/10.48550/arXiv.1810.04805.

Chordia, V. (2021). PunKtuator: A multilingual punctuation restoration system for spoken and written text. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pages 312–320. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2021.eacl-demos.37.

Internet Language Reference Book. (2008–2023). Praha: Ústav pro jazyk český AV ČR. Accessible at: https://prirucka.ujc.cas.cz/.

Karlík, P. (2017). Vokativ. In M. Nekula et al. (eds.): Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/search?action=listpub&search=vokativ.

Kovář, V. et al. (2016). Evaluation and improvements in punctuation detection for Czech. In P. Sojka et al. (eds.): Text, Speech, and Dialogue, pages 287–294. Springer International Publishing.

Lehečka, J. et al. (2021). Comparison of Czech Transformers on Text Classification Tasks. In L. Espinosa-Anke et al. (eds): Statistical Language and Speech Processing. SLSP 2021. Lecture Notes in Computer Science, vol. 13062. Springer. Accessible at: https://doi.org/10.1007/978-3-030-89579-2_3.

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv. /abs/1907.11692. Accessible at: https://doi.org/10.48550/arXiv.1907.11692.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts: Comparative Study. In P. Sojka et al. (eds): Text, Speech, and Dialogue. TSD 2022. Lecture Notes in Computer Science, vol 13502. Springer. Accessible at: https://doi.org/10.1007/978-3-031-16270-1_10.

Nunberg, G. (1990). The Linguistics of Punctuation. CSLI lecture notes. Cambridge University Press.

Radford, A. et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2018, pages 111–123.

Švec, J. et al. (2014) General framework for mining, processing and storing large amounts of electronic texts for language modelling purposes. Lang Resources & Evaluation 48, pages 227–248. Accessible at: https://doi.org/10.1007/s10579-013-9246-z.

Švec, J. et al. (2021). Transformer-based automatic punctuation prediction and word casing reconstruction of the ASR output. In K. Ekštein et al. (eds.): Text, Speech, and Dialogue, Springer International Publishing, pages 86–94.