Study on Automatic Punctuation Restoration in Bilingual Broadcast Stream

Martin Polacek
AILab@TUL
FM TUL
Liberec, Czech Republic
martin.polacek@tul.cz

Abstract

In this study, we employ various ELECTRA-Small models that are pre-trained and fine-tuned on specific sets of languages for automatic punctuation restoration (APR) in automatically transcribed TV and radio shows, which contain conversations in two closely related languages. Our evaluation data specifically concerns bilingual interviews in Czech and Slovak and data containing speeches in Swedish and Norwegian. We train and evaluate three types of models: the multilingual (mELECTRA) model, which is pre-trained for 13 European languages; two bilingual models, each pre-trained for one language pair; and four monolingual models, each pre-trained for a single language. Our experimental results show that a) fine-tuning, which must be performed using data belonging to both target languages, is the key step in developing a bilingual APR system and b) the mELECTRA model yields competitive results, making it a viable option for bilingual APR and other multilingual applications. Thus, we publicly release our pre-trained bilingual and, in particular, multilingual ELECTRA-Small models on HuggingFace, fostering further research in various multilingual tasks.

1 Introduction

In recent years, multiple automatic speech recognition (ASR) systems have been developed, becoming integral to our daily interactions with technology. These systems are now widely used in virtual assistants, automated transcription tools, and numerous other applications that convert spoken language into text.

A key factor driving this widespread adoption has been the development of advanced deep learning methods, particularly end-to-end (E2E) systems (Li, 2022). Unlike traditional ASR approaches that require separate stages for feature extraction, acoustic modeling, and language modeling, E2E systems adopt a more streamlined architecture. They directly map audio inputs to textual outputs, reducing complexity and often improving accuracy. This breakthrough has enabled the creation of ASR models for many languages (Toshniwal et al., 2017), broadening worldwide access to speech recognition technology.

Despite these advancements, some ASR systems still face significant challenges, one of the most notable being the absence of punctuation marks in their output. This limitation arises primarily from the nature of training data, which sometimes does not include punctuation information. Consequently, some ASR models then produce a continuous stream of words without the linguistic boundaries necessary for clear and structured text.

The lack of punctuation negatively impacts both user experience and downstream tasks. For example, in live captioning scenarios, the absence of sentence boundaries can make text difficult to read and understand, particularly in fast-paced or complex dialogues. To solve this issue, modules for automatic punctuation restoration (APR) are usually employed at the output of many ASR systems. In most cases, however, these modules are pre-trained for only one target language, preventing them from correctly formatting the output of multilingual ASR systems (Li et al., 2022) that can process data streams containing utterances in more than one language.

2 Motivation for this work

This work focuses on a specific task of APR in transcribed TV/R (TV and radio) streams containing speech in two similar languages. This phenomenon occurs often in neighboring countries (regions) where people speak a similar or mutually intelligible language.

Proceedings of the Student Research Workshop associated with RANLP-2025,
pages 37–43, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/issn.2603-2821.2025_005

For example, the Czech and Slovak Republics formed one state, Czechoslovakia, between 1918 and 1992; many people born in one country now live in the second one. The two languages are thus similar in that native speakers of Czech understand Slovak and vice versa. The situation is similar in Scandinavia. Here, the population speaks many related languages and dialects, the most widespread of which is the triplet comprising Swedish, Danish, and Norwegian. Norwegian has many similarities with the first two languages, so a native Norwegian speaker can understand Danish and Swedish. Therefore, a Norwegian TV program may often feature a person speaking Swedish or Danish. A third example is the former Yugoslavia, which includes mainly Croatia, Serbia, Bosnia and Herzegovina, and Montenegro. The people living here speak mutually intelligible languages belonging to the western branch of the South Slavic languages.

TV/R programs in these regions containing speech in more than one language are often bilingual. These are typically interviews or talk shows in which the invited person or presenter speaks a different language from the invited guest. In the Czech Republic, for example, there are many interviews with Slovak guests on the Czech television station DVTV. On Slovak television, on the other hand, many Czech guests have appeared on the talk show "Trochu inak s Adelou". Another example is the popular talk show Skavlan, broadcast on Norwegian, Swedish, and Danish television between 2009 and 2021. The Norwegian presenter Fredrik Skavlan invited various speakers of different Scandinavian languages to the show.

Finally, it should be noted that the issue of transcription and APR in bilingual streams also relates to the task of live subtitling of various conferences or social events. For example, it is common for a conference held in the Czech Republic to feature speakers of Czech and Slovak and vice versa.

3 Related work and our contribution

The first developed APR methods were purely statistical. Their biggest drawbacks were their heavy dependence on the quality of the ASR output and low robustness to words outside the system dictionary. The latter problem is becoming increasingly acute with the shift from dictionary-based ASR models to end-to-end (E2E) systems.

In the next phase of development, recurrent neural networks (RNNs) have begun to be used (Kim, 2019), which have shown significantly better performance and allowed the incorporation of both textual and prosodic features. However, their use poses a challenge, especially regarding efficient training data preparation. With the advent of attention mechanism-based transformers, the BERT architecture (Devlin et al., 2019) was among the first used for APR. It outperforms the models with LSTM (Hochreiter and Schmidhuber, 1997) and BiLSTM layers (Tilk and Alumae, 2015) by more than 30% (Polacek et al., 2023). All the previously mentioned models were pre-trained (and fine-tuned) for only one language. However, in 2019, multiple languages were combined during pre-training to create the M-BERT (Pires et al., 2019) model, which can understand the word-to-word connections between languages and thus works for, e.g., the speech translation task.

In this work, we take advantage of the ELECTRA-Small architecture, which achieved better results for the APR task than the BERT model in our previous study (Polacek et al., 2023); we train one multilingual model for 13 selected European languages and two bilingual models Czech-Slovak (CZ+SK) and Norwegian-Swedish (NO+SE) to enhance language modeling across closely related languages by leveraging shared linguistic features. We then investigate the performance of all these models fine-tuned for APR on bilingual corpora and compare their results to monolingual models that are fine-tuned on the same data. For this evaluation, we utilize a dataset containing bilingual Czech-Slovak and Norwegian-Swedish texts. Our results show that pre-training and fine-tuning on data belonging to both languages are necessary to achieve the best performance and that the difference between the results of monolingual and bilingual models increases with the bigger the language difference. We also make public pre-trained multilingual and bilingual models on the HuggingFace platform (Wolf et al., 2020) under a CC-BY-4.0 license:

- Czech-Slovak bilingual model [1].
- Norwegian-Swedish bilingual model [2]
- Multilingual model (mELECTRA) for main European languages [3]

[1] https://huggingface.co/AILabTUL/BiELECTRA-czech-slovak

[2] https://huggingface.co/AILabTUL/BiELECTRA-norwegian-swedish

[3] https://huggingface.co/AILabTUL/mELECTRA

In particular, we believe that a multilingual model can be useful for the community since no one has yet made the ELECTRA architecture in a multilingual version public. Its training requires a lot of data and significant computational resources.

4 Adopted bilingual & multilingual models

As aforementioned, based on our previous research in (Polacek et al., 2023), where we investigated multiple transformer-encoder model types (Vaswani et al., 2017) for the APR task, the neural network architecture adopted in this work corresponds to the ELECTRA-Small model (Clark et al., 2020). This model is complemented with a classification head consisting of one feed-forward layer and one linear layer. The feed-forward layer takes a feature vector of size 256 on input and produces a feature vector of size 512 after passing through the SELU(Klambauer et al., 2017) activation function. The second linear layer produces a probability score for 4 classes (none, dot, comma, and question mark).

4.1 Tokenization

For tokenization of all models, we employ the SentencePiece tokenizer (Kudo and Richardson, 2018) with a vocabulary size of 30,525 as the authors in (Polacek et al., 2023). The amount of data used to create the vocabulary is the same for each language and corresponds to the smallest amount available, i.e., 3.75 GB for Portuguese (the size of training data for each language is summarized in Table 1). This approach ensures an equal data distribution for all languages, mitigating token imbalance. Note that all punctuation marks (e.g., period, question mark, and comma) and numbers are defined as separate tokens during tokenization.

The numbers in Table 2 and Table 3 for Czech and Swedish, respectively, then underscore the efficiency of leveraging language similarity when constructing tokenizers. For example, mixing Slovak and Czech data to create single tokenizer just slightly increases the total token count for the Czech text corpus by 1.2%. Similarly, mixing Norwegian and Swedish data to create a single tokenizer increases the token count on Swedish data by 5.6%, which shows that Swedish and Norwegian are slightly more distant languages than Czech and Slovak.

At the same time, the multilingual tokenizer, which supports 13 languages, extends the token count by just 32.1% on Czech data and by 33.5% on Swedish data. These numbers suggest that the multilingual tokenizer maintains a reasonable balance between flexibility for multiple languages and efficiency for individual ones, making it well-suited for multilingual applications.

4.2 Pre-training and data used

We followed the pre-training procedure outlined for the ELECTRA model in (Clark et al., 2020), adapting it for our multilingual setup. As mentioned, the multilingual model was trained using data from 13 languages, with their representation determined by the availability of language-specific data to ensure a fair and meaningful comparison with single-language models. This approach allowed us to balance the model's capacity to generalize across multiple languages while preserving its performance for individual languages.

The primary data source for each language was transcriptions from TV/R broadcasts, which provide a rich and diverse representation of spoken language. As a complement, smaller portions of the dataset included newspaper articles and legal texts. This data added variety and improved the model's understanding of different text domains. Details of the pre-training data are shown in Table 1. Note the total dataset size was 34.8 GB for the Czech-Slovak (CZ+SK) model, 13.26 GB for the Norwegian-Swedish (NO+SE) model, and 131.51 GB for the multilingual model.

Data preparation for pre-training involved careful processing to ensure consistency. First, all tokenized samples (each containing sentences from a single language) were combined into a unified dataset. These samples were then shuffled across languages to ensure balanced representation within batches. This batching strategy facilitated the model's exposure to diverse linguistic patterns during pre-training, helping it learn shared representations.

For both pre-training and fine-tuning, we used a single NVIDIA H100 GPU (80 GB VRAM) and 16 GB of system RAM. The pre-training of each ELECTRA-Small model took approximately 40 hours. Fine-tuning took about 6 hours per model. All training was performed with mixed-precision (FP16).

Language	SE	SL	SK	PT	PL	NO	IT	HR	FR	EN	DK	DE	CZ
Size [GB]	4.29	6.82	10.9	3.75	11.1	8.97	9.41	14.2	15.7	7.24	4.33	10.9	23.9

Table 1: Summary of training data available for individual languages

Tokenizer	Number of tokens [%]
CZ	100.0
CZ+SK	101.2
Multilingual	132.1

Table 2: Percentage of tokens for Czech data using different tokenizers; CZ tokenizer is the 100% baseline.

Tokenizer	Number of tokens [%]
SE	100.0
NO+SE	105.6
Multilingual	133.5

Table 3: Percentage of tokens for Swedish data using different tokenizers; SE tokenizer is the 100% baseline.

5 Experimental results

5.1 Evaluation metrics

To evaluate the model performance, we used the F1 metric, a commonly used measure combining precision and recall (Van Rijsbergen, 1979). This metric was specifically calculated for classes representing punctuation marks, such as commas, periods, and question marks, as these are critical for assessing the model's ability to predict punctuation correctly. The evaluation utilized a weighted average approach, where the contribution of each class to the final F1 score was proportional to its frequency in the dataset. This ensures that more frequent punctuation marks, which have a greater influence on the overall performance, also have a greater impact on the final score.

To prevent distortion of the results, the "None" class, representing the absence of punctuation, was excluded from the evaluation. Since this class is typically dominant in the dataset, it would disproportionately inflate the F1 score, masking the model's true ability to predict punctuation marks accurately.

5.2 Data used for evaluation

Our data for CZ/SK evaluation consists of manually corrected transcripts of monolingual Czech TV/R and Slovak TV/R news, bilingual interviews with Czech moderators, and Slovak guests (from station DVTV) and bilingual interviews with Slovak moderators, and Czech guests (from the talk show "Trochu inak s Adelou"). This bilingual set is publicly available[4],

For NO/SE evaluation, we also used monolingual Swedish and Norwegian TV/R news transcripts. However, suitable transcriptions for the bilingual scenario were unavailable, as those of bilingual shows exist only in a variant translated to one of the languages. To overcome this issue, we created a synthetic dataset simulating bilingual interviews on various topics: we utilized OpenAI's GPT-4o model (OpenAI, 2023) to generate artificial conversations. First, multiple interview topics were selected, including sports, weather, traveling, culture, gastronomy, hobbies, cooperation, etc. Their generation was then initiated using the prompt:

> "Generate a conversation where one paragraph is in Norwegian and the other paragraph is in Swedish and alternate like this. Write only the paragraphs and generate a long interview. The topic is [TOPIC]"

In total, we created 156 artificial interviews. The resulting bilingual set is also made public[5].

5.3 Effect of data size for fine-tuning

The first performed experiment investigates the effect of the amount of data used for fine-tuning. We fine-tuned the model for APR using the methodology described in our previous work (Polacek et al., 2023), with the only difference in the amount of data used. The results for the Czech ELECTRA model are summarized in Table 4. They show that only 100 MB of data is sufficient for achieving optimal performance. Note that this experiment was performed on the Czech part of the development set described in Section 5.2.

[4] https://owncloud.cesnet.cz/index.php/s/HHfTnWK8D3202Q2
[5] https://owncloud.cesnet.cz/index.php/s/WzqYFR0e1HWbJ66

Table 4: Results of APR after using datasets of various sizes for fine-tuning

# tokens (data size)	P[%]	R[%]	F1[%]
4.2M (25 MB)	74.0	68.8	71.3
8.4M (50 MB)	74.3	69.1	71.6
12.6M (75 MB)	75.3	73.1	74.2
16.8M (100 MB)	**76.4**	**75.5**	**75.9**
33.6M (200 MB)	74.8	74.3	74.5
50.4M (300 MB)	74.6	75.0	74.8
67.2M (400 MB)	75.6	75.5	75.5
84.0M (500 MB)	74.7	74.6	74.6
168.0M (1000 MB)	75.0	74.2	74.6
252.0M (1500 MB)	75.4	75.2	75.3
336.0M (2000 MB)	74.4	74.4	74.4
420.0M (2500 MB)	74.9	74.9	74.9

For bilingual and multilingual models, we selected 100 MB of text data for each language (i.e., 200 MB in total for each bilingual model). To mimic realistic scenarios such as bilingual debates, we split the fine-tuning corpus into individual sentences and randomly mixed them into batches. Additionally, we applied a preprocessing step where, in 50% of the samples, 1–3 words were removed from the beginning and the end of the sequence. This approach improves training data variability and ensures that not all training sequences represent complete sentences.

5.4 Results on Czech and Slovak data

In Table 5, we report the results of the next performed experiment: first, the SK model was fine-tuned on Slovak data only (a), and the CZ model was fine-tuned on Czech data only (b). Next, both models were fine-tuned on a combined CZ+SK dataset (c,d). Subsequently, the mELECTRA model was fine-tuned on the same CZ+SK dataset (e), and finally, the bilingual CZ+SK model underwent fine-tuning on the same data (f).

The yielded results highlight the importance of using data for fine-tuning from both languages intended for inference. This fact follows from the first and second rows, where fine-tuning for just one language leads to a significant performance drop for the second one. In other words, the SK model (a) shows a 15.7% decrease in F1 score on the CZ evaluation dataset compared to the SK model fine-tuned on CZ+SK data (c), and the CZ model (b) shows a 15.9% decrease in F1 compared to the CZ model fine-tuned on CZ+SK data (d). These results also confirm that fine-tuning on both languages yields a noticeable improvement in performance. Furthermore, the results in Table 5 also show very good performance of the mELECTRA model, which, after fine-tuning on Czech and Slovak data, yields just slightly worse F1 values than the best-performing models pre-trained for a single language.

pre-training	**fine-tun.**	**CZ F1 [%]**	**SK F1 [%]**
(a) SK	SK	59.9	71.2
(b) CZ	CZ	**77.8**	57.0
(c) SK	CZ + SK	75.6	**74.4**
(d) CZ	CZ + SK	76.2	72.9
(e) mELECTRA	CZ + SK	76.4	74.1
(f) CZ + SK	CZ + SK	76.0	73.1

Table 5: Comparison of various APR models on monolingual Czech and Slovak datasets

The next experiment, see Table 6, presents the results yielded on bilingual transcripts of TV/R interviews. The obtained results reveal that the performance of all evaluated models is similar, emphasizing the key role of fine-tuning, which enables the models to adapt effectively to the specifics of the target dataset and to mitigate differences arising from pre-training. Notably, the mELECTRA model (e), despite being pre-trained for many languages, performs only 0.5% worse than the best model (f). This small gap demonstrates, similarly to the previous experiment, the potential of a general-purpose multilingual model, which can eliminate the need for pre-training language-specific models for the APR task.

pre-training	**fine-tun.**	**CZ+SK F1 [%]**
(a) SK	SK	62.9
(b) CZ	CZ	61.2
(c) SK	CZ + SK	66.6
(d) CZ	CZ + SK	66.3
(e) mELECTRA	CZ + SK	66.2
(f) **CZ + SK**	**CZ + SK**	**66.7**

Table 6: Comparison of various models on a bilingual Czech/Slovak dataset

5.5 Results on Norwegian and Swedish data

Similar to the previous evaluations, we conducted experiments using models pre-trained on a single language (Norwegian or Swedish), on both languages (NO+SE), and using all available multilingual data (mELECTRA). From Table 7, it is obvious that the performance on the language, which

was not used for fine-tuning, drops by 44.6% for model (a) on Norwegian data and by 47.3% for model (b) on Swedish data. On the contrary, the use of both languages for fine-tuning yields significant improvements in performance for both languages. Specifically, model (c) has an F1 of 1.8% higher on Norwegian data compared to the single-language model (a). Similarly, under the same conditions, model (d) achieved a 0.5% increase in F1 over model (b). The mELECTRA model (e) performed comparably to model (d) on Norwegian data, while its accuracy on Swedish data was only 2.8% lower. The best-performing model (f) was found to be only 2.2% less accurate than model (c) on Norwegian data and 0.9% less accurate than model (d) on Swedish data, demonstrating its competitive performance across both languages.

pre-training	fine-tun.	NO F1 [%]	SE F1 [%]
(a) NO	NO	71.1	19.9
(b) SE	SE	23.8	64.5
(c) NO	NO + SE	**72.9**	52.3
(d) SE	NO + SE	66.6	**65.0**
(e) mELECTRA	NO + SE	67.0	62.2
(f) NO + SE	NO + SE	70.7	64.1

Table 7: Comparison of APR models on monolingual NO and SE datasets

The last experiment presented in Table 8 shows the results for the artificially created bilingual Norwegian/Swedish dataset. Here, it is evident that model (c) achieved the worst results and models (d) and (e) yielded (as expected from the previous experiment) similar F1 values of 71.6% and 71.7%, respectively. The best results were obtained by using model (f). However, its F1 value of 74.2% is just by 2.5% higher than that of the mELECTRA model (e). This means that the model pre-trained on all multilingual data proves good performance not only for the Czech/Slovak data but also for Norwegian/Swedish bilingual interviews.

pre-training	fine-tun.	NO+SE F1 [%]
(a) NO	NO	67.5
(b) SE	SE	64.1
(c) NO	NO + SE	67.8
(d) SE	NO + SE	71.6
(e) mELECTRA	NO + SE	71.7
(f) **NO + SE**	**NO + SE**	**74.2**

Table 8: Comparison of various models on a bilingual Norwegian/Swedish dataset

6 Conclusion

This study demonstrates the necessity of fine-tuning bilingual and multilingual models for APR in bilingual ASR outputs. Our findings indicate that as the difference between the two languages increases, the need for fine-tuning using data belonging to both of them becomes crucial. Without adequate fine-tuning, performance for the untrained language drops significantly.

At the same time, we show that the more the languages differ from each other, the more important the data for pre-training is. For the Czech/Slovak data, there was almost no improvement with pre-training on both languages, whereas, for the Norwegian/Swedish pair, there was already a more than 2% improvement. In other words, using bilingual data only for fine-tuning does not guarantee the best results.

Furthermore, in specific cases such as Slovak, fine-tuning of monolingual, bilingual, and multilingual models on CZ+SK datasets resulted in performance improvements compared to training solely on SK data. This suggests that leveraging linguistic similarities between closely related languages can enhance model robustness and effectiveness beyond single-language training.

Lastly, our study identifies a promising alternative in the use of a pre-trained multilingual model. This type of model can achieve competitive performance with only 100MB of fine-tuning data. This efficiency makes multilingual models an attractive solution for handling APR in bilingual as well as monolingual streams.

7 Acknowledgements

This work was supported by the Student Grant Scheme at the Technical University of Liberec through project nr. SGS-2024-3425. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Seokhwan Kim. 2019. Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In *ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7280–7284. IEEE.

Gunter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In *NIPS 2017, Long Beach, CA, USA, December 4-9, 2017*, pages 971–980.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. ACL.

Bo Li, Tara N. Sainath, Ruoming Pang, Shuo yiin Chang, Qiumin Xu, Trevor Strohman, Vince Chen, Qiao Liang, Heguang Liu, Yanzhang He, Parisa Haghani, and Sameer Bidichandani. 2022. A language agnostic multilingual streaming on-device asr system. In *Interspeech*.

Jinyu Li. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1):1–64.

OpenAI. 2023. Gpt-4 technical report.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics.

Martin Polacek, Petr Cerva, Jindrich Zdánský, and Lenka Weingartová. 2023. Online punctuation restoration using electra model for streaming asr systems. *INTERSPEECH 2023*.

Ottokar Tilk and Tanel Alumae. 2015. LSTM for punctuation restoration in speech transcripts. In *Interspeech 2015, Dresden, Germany, September 6-10, 2015*, pages 683–687. ISCA.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro J. Moreno, Eugene Weinstein, and Kanishka Rao. 2017. Multilingual speech recognition with a single end-to-end model. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.

C. J. Van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworths, London.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, and et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.