From Conversational Speech to Readable Text: Post-Processing Noisy Transcripts in a Low-Resource Setting

Arturs Znotins
IMCS, University of Latvia
arturs.znotins@lumii.lv

Normunds Gruzitis
University of Latvia
normunds.gruzitis@lu.lv

Roberts Dargis
RGP, Latvia
Viroling Technology, Latvia

Abstract

We present ongoing research on automatic post-processing approaches to enhance the readability of noisy speech transcripts in low-resource languages, with a focus on conversational speech in Latvian. We compare transformer-based sequence-labeling models and large language models (LLMs) for the standard punctuation and capitalization restoration task, while also considering automatic correction of mispronounced words and disfluency, and partial inverse text normalization. Our results show that very small LLMs (approx. 2B parameters), fine-tuned on a modest text corpus, can achieve near state-of-the-art performance, rivaling orders of magnitude larger LLMs. Additionally, we demonstrate that a fine-tuned Whisper model, leveraging acoustic cues, outperforms text-only systems on challenging conversational data, even for a low-resource language. Error analysis reveals recurring pitfalls in sentence boundary determination and disfluency handling, emphasizing the importance of consistent annotation and domain adaptation for robust post-processing. Our findings highlight the feasibility of developing efficient post-processing solutions that significantly refine ASR output in low-resource settings, while opening new possibilities for editing and formatting speech transcripts beyond mere restoration of punctuation and capitalization.

1 Introduction

Automatic punctuation and capitalization restoration has been widely studied as a post-processing step for automatic speech recognition (ASR) systems, aiming to improve transcript readability and facilitating downstream NLP tasks such as machine translation, named entity recognition, etc.

Early methods leveraged statistical approaches, such as n-gram language modeling and prosodic cues (Stolcke et al., 1998; Beeferman et al., 1998), as well as sequence labeling techniques like Conditional Random Fields (CRFs) (Lu and Ng, 2010; Wang et al., 2012) and Maximum Entropy models. With the advent of deep learning, recurrent neural networks (RNNs) and long short-term memory (LSTM) models proved to be more efficient in modeling sequential dependencies (Tilk and Alumäe, 2015). Bidirectional RNNs and transformer-based architectures further enhanced accuracy by using richer contextual representations (Yi and Tao, 2019; Nguyen and Salazar, 2019).

Recent work has demonstrated that transformer-based models outperform previous neural approaches. BERT-based models, such as RoBERTa and ELECTRA, have achieved state-of-the-art results on punctuation restoration by leveraging large-scale pretraining (Devlin et al., 2018; Poláček et al., 2023). Other studies have explored multilingual transformer models such as XLM-RoBERTa (Conneau et al., 2020) to address punctuation restoration across multiple languages.

End-to-end ASR models, such as OpenAI's Whisper (Radford et al., 2023), directly generate transcriptions with punctuation and capitalization. Whisper is trained on large-scale weakly supervised data, allowing it to outperform conventional ASR models that require separate punctuation restoration modules.

Recent advances in large-scale auto-regressive large language models (LLMs), such as GPT-4 (OpenAI et al., 2024), have introduced new paradigms for punctuation restoration. Unlike conventional sequence labeling approaches, GPT-style models perform text infilling and editing, enabling them to restore punctuation in a generative manner. Recent developments in open-source multilingual LLMs have led to the creation of smaller models that effectively support low-resource languages (Dargis et al., 2024).

Developing robust punctuation restoration models relies on sufficiently large and representative annotated corpora. Europarl (Koehn, 2005) and TED-LIUM (Rousseau et al., 2014) have been widely

used, but they often lack domain-specific noise typical of real-world ASR. Fu et al. (2021) demonstrated that domain-adaptive fine-tuning with n-gram similarity-based data sampling can improve model robustness. Data augmentation methods that simulate ASR errors have also been shown to yield significant performance gains (Alam et al., 2020).

For Latvian (approx. 1.5M native speakers), relevant work on punctuation restoration includes (Salimbajevs, 2016; Vāravs and Salimbajevs, 2018), which focus on bidirectional models and sequence labeling for punctuation and capitalization. One publicly available resource is a proprietary on-line service that allows users to correct the punctuation and formatting of a text, where the underlying model, likely an encoder-decoder, is trained on academic texts[1]. Another publicly available resource is an open-source punctuation model based on XLM-RoBERTa[2] (Guhr et al., 2021), trained on Europarl data. The best available end-to-end Latvian ASR models that include text formatting are *whisper-large-v3* and *whisper-large-v3-lv*, the latter being fine-tuned on the dataset described in Section 2 as well as on the Common Voice 19.0 dataset[3] (Dargis et al., 2024).

Our contributions in this study are as follows:

- We demonstrate that even the smallest generative LLMs (i.e., in the 2B parameter range) can be fine-tuned on a relatively small text corpus to achieve near state-of-the-art results, bridging the gap between reference text formatting and noisy ASR output.
- We present a thorough error analysis highlighting common pitfalls, such as mispunctuation, ambiguous sentence boundaries, and speaker disfluencies.
- Beyond punctuation and capitalization, we show that LLMs can partially learn error correction and inverse text normalization from limited data, underlining their potential to further refine ASR outputs in low-resource settings.

[1] https://salieckomatus.lv

[2] https://huggingface.co/1-800-BAD-CODE/xlm-roberta_punctuation_fullstop_truecase

[3] https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-lv-late-cv19

2 Dataset

We use the LATE-Media corpus[4] (Auzina et al., 2024a,c), which comprises approximately 70 hours of conversational Latvian speech from broadcast recordings, sourced from public media. The data includes both spontaneous and prepared speech (but not read speech) from more than 250 speakers, offering a diverse range of speaking styles and topics.

Transcriptions are provided in standard Latvian orthography, with additional punctuation and grammar rules applied. When necessary, annotations in square brackets capture non-standard pronunciation (e.g., "lasām [lasam]") and foreign words (e.g., "Rail [reil] Baltica [boltik]"). The corpus also documents the reading of numbers, accounting for syntactic agreement in context (e.g., *nominative* vs. *dative* forms). This rich annotation scheme ensures that spontaneous variations – such as word repetitions, truncated words, and different realizations of abbreviations – are properly represented.

To simplify the punctuation restoration task, we unify several less frequent or inconsistently annotated marks by replacing them with periods. Specifically, we map exclamation marks, ellipses, and em dashes to periods. We also ignore seldom-used marks such as colons and semicolons, which tend to be subjectively annotated. These steps reduce annotation noise and help stabilize model performance in subsequent training and evaluation.

The dataset statistics, including the distribution of punctuation, capitalization types, average sentence length, and correction annotations, are presented in Table 1.

	Train	Dev	Test
Comma	56268	1624	1595
Period	49599	2363	2389
Question	3454	409	419
Title	73675	3172	3257
Upper	3528	87	78
Avg Sent Len	10.3	7.6	7.2
Corrections	4541	80	85

Table 1: Dataset statistics for punctuation, capitalization and sentence lengths.

[4] https://korpuss.lv/en/id/LATE-mediji

3 Experimental Setup

In our experiments, we address the following key research questions:

- How do small generative models compare to larger models in punctuation and capitalization restoration tasks for a low resource language, and how does their performance degrade on ASR-generated transcripts?
- To what extent are models capable of correcting transcript text without introducing unnecessary modifications?
- What are the predominant error types?

We evaluate two distinct scenarios: formatting ASR-generated transcripts and formatting manually transcribed reference text. This setup allows us to assess how models handle noisy ASR outputs and whether they can refine reference transcripts without unnecessary modifications. The evaluation of ASR-generated transcripts is conducted on the outputs of *whisper-large-v3-lv*, currently the strongest open-source Latvian ASR model. We use publicly available *salieckomatus.lv* and XLM-RoBERTa (Guhr et al., 2021) as baselines. Additionally, we evaluate the performance of *whisper-large-v3* and *whisper-large-v3-lv*.

Performance is measured using F1-score (F1) for punctuation restoration and capitalization. To ensure that models do not introduce unnecessary modifications, we also compute the word error rate (WER) on the normalized formatted transcript. A heuristic fuzzy alignment method is used to align incorrectly recognized words and words that differ in spoken and written forms, such as number expressions, acronyms, and abbreviations.

For LLMs, we employ the following task-specific prompt:

> *"You are a skilled editor specializing in Latvian transcripts. Your task is to format this short (under 30 seconds) ASR-produced transcript by adding punctuation (use only commas, periods, and question marks), capitalization, and making minimal edits for readability. Correct grammar, mispronounced words, and abbreviations as needed. Convert numbers into their written form. Do not alter the sentence structure or meaning – only refine specific words, punctuation, and formatting while keeping it as close to the original as possible."*

For fine-tuned models, we use a shorter prompt, observing no noticeable drop in performance:

> *"Proofread the provided Latvian transcript by inserting appropriate punctuation and applying proper capitalization."*

Models are fine-tuned exclusively on the training split, without incorporating any external data. The fine-tuning uses a linearly decreasing learning rate of 2e-5, a warm-up ratio of 0.1, a batch size of 32, and runs for 3 epochs.

4 Results

The results of our experiments are presented in Table 2. Generative models, such as GPT-4o and GPT-4o *mini*, demonstrate strong capabilities for Latvian punctuation and capitalization tasks. However, they also introduce unintended transcript modifications, reflected in elevated WER – an issue which can potentially be mitigated with more extensive prompt optimization.

Fine-tuned (FT) models show significant gains in consistency, with GPT-4o FT achieving the highest overall performance (F1 scores of 81.5 for punctuation and 84.4 for capitalization). Notably, smaller fine-tuned models (e.g., Gemma-2B, EuroLLM-1.7B) perform at levels comparable to GPT-4o *mini*, suggesting that the model size alone does not dictate effectiveness for this task.

Table 3 highlights a key limitation of generative models if compared to BERT-based models – unintended alterations to the transcripts. This issue is especially pronounced in the case of non-fine-tuned models. GPT-4o, for example, often attempts to enhance fluency by removing words deemed superfluous (e.g., "*And my*" → "*My*") or by adding implied speech elements (e.g., "*tea, coffee*" → "*tea or coffee*"). The most frequently observed and potentially the most influential errors are word substitutions that alter the meaning or introduce syntactic agreement errors. Although prompt optimization can partially address these issues, they remain challenging to be completely eliminated without highly descriptive prompting and provision of examples for in-context learning.

Smaller generative models like GPT-4o *mini* can introduce more pronounced substitutions (e.g., "*bračka*" (brother) → "*brāķa*" (defect)) as well as occasionally produce non-existent words (e.g.,

Model	Punctuation				Capitalization			WER
	Comma	Period	Question	Total	Title	Upper	Total	
whisper-large-v3	64.1	73.0	63.3	68.4	72.2	41.7	72.0	31.3
whisper-large-v3-lv	77.5	79.9	72.1	78.3	81.9	53.3	81.8	12.7
ASR Output								
XLM-RoBERTa	74.7	78.8	57.5	75.1	79.0	46.2	78.9	12.7
salieckomatus.lv	74.9	75.9	40.7	73.1	77.6	22.2	77.4	13.1
GPT-4o	78.1	80.8	62.6	78.2	81.4	36.4	81.3	13.2
GPT-4o FT	**81.2**	**84.0**	**68.9**	**81.5**	**84.6**	38.5	**84.4**	**12.5**
GPT-4o mini	74.4	80.0	57.5	75.8	79.8	36.4	79.7	15.2
GPT-4o mini FT	79.3	82.0	64.2	79.3	82.5	41.7	82.4	12.7
EuroLLM-1.7B-Instruct FT	78.8	82.5	60.1	79.0	82.8	**53.8**	82.7	12.8
gemma-2-2b-it FT	79.5	81.5	63.2	79.1	82.0	41.7	81.9	12.7
Reference Transcripts								
XLM-RoBERTa	77.9	79.7	62.4	77.4	84.3	72.7	84.3	0.0
salieckomatus.lv	77.6	76.7	43.2	74.8	81.6	43.5	81.4	1.4
GPT-4o	80.9	82.4	67.6	80.5	86.6	48.3	86.4	3.0
GPT-4o FT	**85.9**	**86.0**	**76.5**	**85.1**	**91.8**	83.9	**91.8**	0.4
GPT-4o mini	76.7	81.5	63.2	78.0	84.9	58.3	84.8	4.7
GPT-4o mini FT	83.1	84.0	70.0	82.4	89.7	80.0	89.6	0.3
EuroLLM-1.7B-Instruct FT	83.2	84.9	69.1	82.8	89.6	**90.3**	89.6	0.4
gemma-2-2b-it FT	82.7	83.1	68.1	81.6	88.3	75.0	88.3	0.4

Table 2: Results on test split: F1 scores for punctuation and capitalization, and WER.

"*paliec*" (stay) → "*paliķ*" (∅), "*filmēs*" (will shoot) → "*filmes*" (∅)).

Overall, fine-tuning reduces unintended text changes by an order of magnitude for all model sizes. While fine-tuned models in the two billion parameter range rarely alter transcripts, the errors they produce typically manifest as ungrammatical forms rather than semantic substitutions.

The Whisper model fine-tuned for Latvian (i.e., *whisper-large-v3-lv*) achieves WER of 12.7, significantly outperforming the base Whisper *large-v3* model while maintaining strong punctuation and capitalization scores.

Models generally perform better on reference transcripts than on ASR outputs, which is expected since ASR-generated text contains recognition errors that interfere with punctuation and capitalization. Similarly, fine-tuned LLMs outperform their non-fine-tuned counterparts when applied to ASR outputs.

We manually annotated 100 samples to analyze errors made by the various models. In the cases of mismatched predictions, we categorized errors as follows:

- Actual errors: incorrect punctuation placement, capitalization mistakes, or misinterpretation of sentence boundaries (57 cases).
- Alternative formatting choices: instances where a model's output differs from the reference but remains grammatically valid (43 cases).

For 21 of the cases, we had to listen to the audio to apply a correct markup, highlighting the importance of audio features, for example, "*Labi. Sakratīts ir.*" ('*Well. Shaken [it] is.*') vs. "*Labi sakratīts ir.*" ('*Well shaken [it] is.*'). This also explains the better question mark performance for *whisper-large-v3-lv* without any extra processing.

We further evaluated model performance in correcting mispronounced words: by using annotated mispronunciations, number expressions, and generally acceptable written forms annotated in the dataset. Approximately 50% of these cases were correctly replaced by LLMs, suggesting a potential for these models to learn error correction and inverse text normalization tasks for the low-resource Latvian from relatively small datasets. However, further investigation is needed, since the current test set is too small for a reliable evaluation.

We have also evaluated a broader set of punctuation marks. However, because of their low fre-

Model	Changed Utt.	Substitute	Inflect	Delete	Insert
XLM-RoBERTa	0.0				
salieckomatus.lv	10.0	58	16	16	11
GPT-4o	16.4	42	21	29	8
GPT-4o FT	**1.4**	79	14	0	7
GPT-4o mini	24.6	41	24	28	7
GPT-4o mini FT	1.5	67	25	8	0
EuroLLM-1.7B-Instruct FT	2.7	69	8	15	8
gemma-2-2b-it FT	1.7	47	27	20	7

Table 3: Error analysis of changed utterances by error type, based on a manual review of a sample of 100 utterances (or fewer if fewer were found) in each model's test split. All values are percentages.

quency beyond commas, periods, and question marks, these results can currently only be considered preliminary and are not yet reliable. Moreover, their usage in conversational ASR transcripts is often subjective, justified by increased inter-annotator disagreement.

5 Conclusion and Further Work

End-to-end ASR systems, such as Whisper fine-tuned for Latvian (using a relatively small amount of data), already provide reasonably well-formatted transcripts for general-domain speech by leveraging acoustic features that are unavailable in text-only approaches. However, even without acoustic cues, formatting performance can be improved with LLMs using a prompt-based approach. Further task-specific fine-tuning yields the best and most stable results, and it is feasible even with smaller LLMs in the 2B parameter range on a small dataset. Larger models often provide higher accuracy but come with increased computational costs and deployment complexity.

Audio features (pauses, intonation) remain a crucial signal for punctuation restoration. Sentence boundaries in speech are often ambiguous, with multiple valid interpretations, and better annotation guidelines could improve consistency. One major challenge in training ASR models for Latvian and other low-resource languages is the lack of datasets that include both conversational speech and formatted transcriptions. LLMs enable transcript transformations such as inverse text normalization and error correction by leveraging their built-in language knowledge, even when trained on relatively small datasets. Thus, fine-tuned LLMs can expedite the addition of such formatting to existing orthographically transcribed datasets, for instance, the LATE-Conversational speech corpus (Auzina et al., 2024b) which comprises 35 hours of informal conversations in Latvian – this is currently a work in progress, to be followed by human verification and evaluation.

6 Limitations

Our models are evaluated on a single dataset for Latvian, limiting generalizability to other domains or languages. Future research should extend these evaluations to multiple datasets.

ASR errors significantly impact formatting performance. Introducing ASR-like noise or synthetic errors during training could improve robustness but risks unintended meaning changes if not done carefully.

In fields like law or medicine, over-corrections can subtly alter meaning. Generative and punctuation models may introduce edits beyond basic formatting, risking inaccuracies in sensitive transcripts. Hence, they should be used cautiously when exact fidelity to the original speech is required.

Acknowledgments

This work was funded by the European Union Recovery and Resilience Facility projects "Language Technology Initiative" (2.3.1.1.i.0/1/22/I/CFLA/002) and "Competence Centre of Information and Communication Technologies" (5.1.1.2.i.0/1/22/A/CFLA/008).

References

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for high- and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT)*, pages 132–142.

Ilze Auzina, Roberts Dargis, Kristine Levane-Petrova, Arta Auzina, Baiba Saulite, Ilze Laksa-Timinska, Elina Gailite, Gunta Nespore-Berzkalne, Guna Rabante-Busa, Kristine Pokratniece, and Agute Klints. 2024a. LATE Media Speech Corpus V1 (LATE-mediji). CLARIN-LV digital library at IMCS, University of Latvia.

Ilze Auzina, Roberts Dargis, Guna Rabante-Busa, Ilze Timinska-Laksa, Elina Gailite, and Arta Auzina. 2024b. LATE Conversational Speech Corpus V1 (LATE-sarunas). CLARIN-LV digital library at IMCS, University of Latvia.

Ilze Auzina, Normunds Gruzitis, Roberts Dargis, Guna Rabante-Busa, Didzis Gosko, Janis Vempers, Raivis Kivkucans, and Arturs Znotins. 2024c. Recent Latvian Speech Corpora for Linguistic Research and Technology Development. *Baltic Journal of Modern Computing*, 12(4):646–658.

Doug Beeferman, Adam Berger, and John Lafferty. 1998. Cyberpunc: A lightweight punctuation annotation system for speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 689–692.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Roberts Dargis, Arturs Znotins, Ilze Auzina, Baiba Saulite, Sanita Reinsone, Raivis Dejus, Antra Klavinska, and Normunds Gruzitis. 2024. BalsuTalka.lv – Boosting the Common Voice Corpus for Low-Resource Languages. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 2080–2085.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan TN, and Simon Corston-Oliver. 2021. Improving punctuation restoration for speech transcripts via external data. *arXiv:2110.00560*.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2021. FullStop: Multilingual Deep Models for Punctuation Prediction. In *Proceedings of the Swiss Text Analytics Conference*. CEUR Workshop Proceedings.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 177–186.

Toan Q Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *arXiv:1910.05895*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, et al. 2024. GPT-4 Technical Report. *arXiv:2303.08774*.

Martin Poláček, Petr Červa, Jindřich Žďánský, and Lenka Weingartová. 2023. Online Punctuation Restoration using ELECTRA Model for streaming ASR Systems. In *Interspeech*, pages 446–450.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518.

Anthony Rousseau, Paul Deléglise, Yannick Esteve, et al. 2014. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *LREC*, pages 3935–3939.

Askars Salimbajevs. 2016. Bidirectional LSTM for automatic punctuation restoration. In *Human Language Technologies – The Baltic Perspective*, pages 59–65. IOS Press.

Andreas Stolcke, Elizabeth Shriberg, Rebecca A Bates, Mari Ostendorf, Dilek Zeynep Hakkani, Madelaine Plauche, Gökhan Tür, and Yu Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *ICSLP*, volume 2, pages 2247–2250.

Ottokar Tilk and Tanel Alumäe. 2015. LSTM for punctuation restoration in speech transcripts. In *Interspeech*, pages 683–687.

Andris Vāravs and Askars Salimbajevs. 2018. Restoring punctuation and capitalization using transformer models. In *6th International Conference on Statistical Language and Speech Processing (SLSP)*, pages 91–102.

Xuancong Wang, Hwee Tou Ng, and Khe Chai Sim. 2012. Dynamic conditional random fields for joint sentence boundary and punctuation prediction. In *Interspeech*, pages 1384–1387.

Jiangyan Yi and Jianhua Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274.