Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced Transcription with RAG-based Correction and Predicted BLEU

Vincenzo Timmel[1], Manfred Vogel[1], Daniel Perruchoud[1], Reza Kakooee[1]

[1]University of Applied Sciences and Arts Northwestern Switzerland

{vincenzo.timmel, manfred.vogel, daniel.perruchoud, reza.kakooee}@fhnw.ch

Abstract

This paper presents a new long-form release of the Swiss Parliaments Corpus, converting entire multi-hour Swiss German debate sessions (each aligned with the official session protocols) into high-quality speech–text pairs. Our pipeline starts by transcribing all session audio into Standard German using Whisper Large-v3 under high-compute settings. We then apply a two-step GPT-4o correction process: first, GPT-4o ingests the raw Whisper output alongside the official protocols to refine misrecognitions, mainly named entities. Second, a separate GPT-4o pass evaluates each refined segment for semantic completeness. We filter out any segments whose Predicted BLEU score (derived from Whisper's average token log-probability) and GPT-4o evaluation score fall below a certain threshold. The final corpus contains 801 hours of audio, of which 751 hours pass our quality control. Compared to the original sentence level SPC release, our long-form dataset achieves a 6-point BLEU improvement, demonstrating the power of combining robust ASR, LLM-based correction, and data-driven filtering for low-resource, domain-specific speech corpora.

1 Introduction

Data scarcity in low-resource domains still hinders the development of Automatic Speech Recognition (ASR) systems. For Swiss German, (Plüss et al., 2021) contributed the Swiss Parliaments Corpus (SPC), including a meticulously prepared training dataset with high alignment quality of 176 hours of Swiss German speech paired with Standard German transcripts of Bernese parliamentary debates with a corresponding curated test dataset of 6 hours. The corpus was built using a forced sentence alignment procedure and alignment quality estimator that overcomes challenges such as sentence reordering and language mismatches between Swiss German audio and Standard German text. They used a global alignment algorithm based on Needleman-Wunsch and an Intersection over Union (IoU) estimator to filter out poor-quality alignments. Additional filters, such as character-per-second limits and language detection, ensured that only accurately aligned sentences were included.

The SPC_R corpus presented in this paper is an extension of the original SPC corpus focusing on the creation, curation, and release of datasets tailored to Swiss German NLP applications. Originally, crawled data from the parliament debates of the Grosser Rat Kanton Bern encompass 801 hours of session recordings in long-form with a length spanning from 28 to 242 minutes paired with official session protocols.

In contrast to (Plüss et al., 2021), which extracts sentences from parliamentary sessions by finding near-perfect matches between automatically generated transcriptions and the official session protocols, we incorporate an advanced transcription pipeline in SPC_R. This includes the Whisper Large-v3 model (Radford et al., 2023) for transcription, and a post-correction step using GPT-4o (Hurst et al., 2024), aligned with the official protocol to further enhance transcription quality and overall data accuracy.

In addition, the SPC_R corpus provides the data in long-form, whereas the original SPC is segmented at sentence level.

The primary contributions include:

- High-quality transcription by Whisper Large-v3 of approximately 801 hours of audio with high-compute settings, see Section 3.
- BLEU score (Papineni et al., 2002) prediction based on Whisper transcription outputs via linear regression.
- A two-step large language model (LLM) approach in which a first model corrects the tran-

scription and a second, independent model evaluates that correction.

This paper provides detailed insights into the methodology, experimental results, and implications for future NLP dataset releases in Swiss German.

2 Related Work

In the past years, several initiatives (Plüss et al., 2021, 2022, 2023; Dogan-Schönberger et al., 2021) made valuable contributions for the development of Swiss German ASR solutions; an overview of the released datasets is shown in Figure 1. However, these datasets are all at sentence level which typically does not improve ASR solutions for real-world situations (Timmel et al., 2024). Additionally, not all existing datasets can be used for commercial purposes.

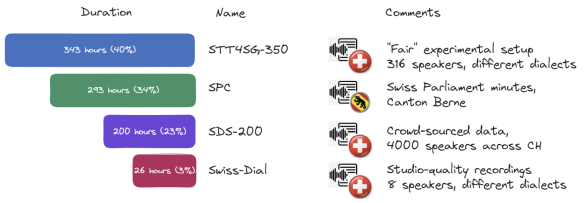

Figure 1: Overview of Swiss German speech to German text datasets. Usage of SPC is possible under MIT license, SDS-200 and STT4SG-350 under SwissNLP license. SwissDial can be used exclusively for research purposes.

3 Transcription with Whisper Large-v3

The starting point for the construction of the SPC_R Corpus is 801 hours of long-form audio from parliament debates of Grosser Rat Kanton Bern which we transcribe with Whisper Large-v3.

Our transcription pipeline uses Whisper Large-v3 via WhisperX (Bain et al., 2023) under high-compute settings, namely *beam_size* set to 10, *best_of* set to 10, and *log_prob_threshold* set to –2. All transcriptions are performed on an NVIDIA A4500 GPU with 20 GB of VRAM, using *float16* precision and a *batch_size* of 8. These high-compute settings further improve results, as shown in Figure 5. For all transcribed parliament sessions, we store Whisper's *avg_log_prob* output, which reflects the model's prediction confidence and exhibits strong predictive power for transcription quality, as described in Subsection 3.1.

3.1 BLEU Prediction

We observed a linear relationship between the confidence metric calculated by Whisper (Kim, 2023), as presented in Equation 1, and the BLEU score (sacreBLEU[1], more precisely) of datasets transcribed with Whisper.

$$\text{confidence} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \quad (1)$$

The confidence is derived from Whisper's segment-specific average log-probabilities *avg_log_prob*, which are averaged over the whole audio file. In Equation (1), p_i denotes the average log-probability for the ith segment, and N is the total number of segments in the entire audio file, where a segment is the text between two timestamps predicted by Whisper. Thus, the confidence is the exponential of the average *avg_log_prob* over a whole audio file.

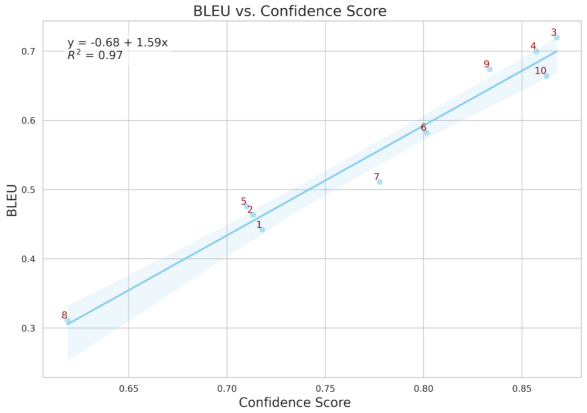

Figure 2: Linear relationship between BLEU score vs. Whisper confidence score for ten long-form conversations, represented by numbers 1-10. The blue shaded area represents the 95% confidence interval.

Figure 2 shows this linear relationship between the BLEU score (calculated between the transcription and a manually created ground truth) and the confidence on ten distinct, independent Swiss German datasets. Each dataset of approximately one hour (ca. 8'000 tokens) consists of manually transcribed Swiss German conversations (the ground truth) between two or more speakers (these datasets cannot be disclosed due to data privacy and NDA restrictions). Our analysis shows that higher confidence values are associated with higher BLEU

[1] https://github.com/mjpost/sacreBLEU (default settings: 4-gram, standard tokenization and smoothing)

scores in a near-linear fashion, indicating that the confidence metric is a strong predictor of transcription quality, suggesting its potential for assessing transcription performance.

A linear regression fitted to these data produced an intercept of -0.68 and a slope coefficient of 1.59 and allows to predict a BLEU score based solely on the confidence, called the Predicted BLEU, without first creating a ground truth.

Figure 3 shows the distribution of Predicted BLEU scores for all 131'291 segments of SPC_R, corresponding to a total of 801 hours of audio.

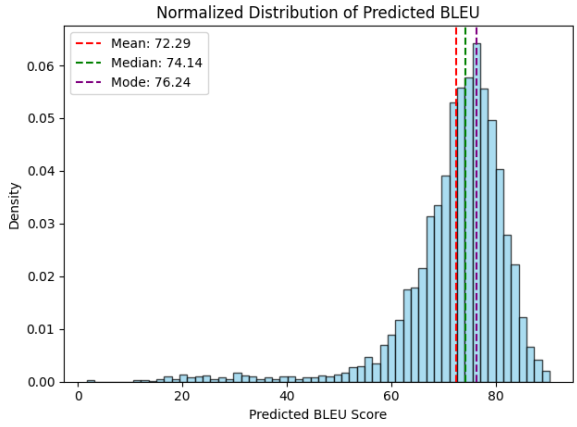

Figure 3: Distribution of Predicted BLEU scores across SPC_R (N = 131'291 data segments).

Figure 4 shows the cumulative proportion of data samples for a given Predicted BLEU score threshold. As the threshold rises, fewer samples qualify, underscoring the balance between transcription quality and the amount of available data.

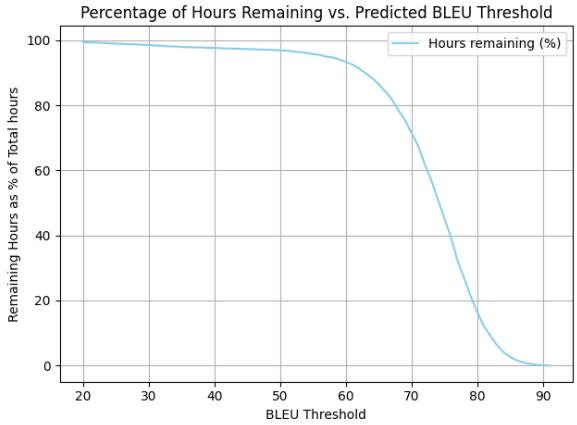

Figure 4: Percentage of data samples that have a BLEU score above the threshold.

Hence, the Predicted BLEU score derived from Whisper's *avg_log_prob* can be used to identify and select high-quality transcription segments (see Section 5).

4 Transcript correction using GPT-4o

Automated transcription with Whisper Large-v3 shows promising results but leads to errors in named entities (e.g., "Alba Rutschi" instead of "Alberucci") and other similar errors. To mitigate this, we introduce a two-step correction process using text-embedding-3-large GPT-4o and GPT-4o-mini (OpenAI, 2023):

1. **Correction Stage:** GPT-4o is used to refine the initial transcription by prompting it to correct errors, segment by segment. Corrections are based on information injected from the official manual summaries of the parliament session corresponding to the audio segment using Retrieval-Augmented Generation (RAG, see Subsection 4.1).

2. **Evaluation Stage:** Evaluation assessments of GPT-4o corrections use manual inspection on small data samples and GPT-4o-mini-as-a-Judge.

GPT-4.1 (OpenAI, 2025) was also evaluated but we found that it would repeatedly change conjugation of words, thus sometimes introducing new errors in the transcription. While still overall reducing the WER, it fixed less errors than GPT-4o.

4.1 Context provision via RAG

RAG (Lewis et al., 2020) is used to provide GPT-4o with factual context to correct the transcription.

We follow best practices (Wang et al., 2024), using Faiss (Douze et al., 2024) for efficient vector storage and retrieval, a sliding window approach and text-embedding-3-large as embedding model. Official manual summaries are ingested with *pyPDF* (Fenniak et al., 2024) using chunks of 600 characters with an overlap of 450. These values are chosen to consistently ensure a complete overlap between the transcription and the context from the chunk based on the maximum segment length of 423 characters. We pass the most relevant chunk to GPT-4o as context without re-ranking retrieved chunks.

Manual evaluation on 122 audio segments corresponding to 50 minutes of transcribed data shows that the correct chunk from the official manual summary is retrieved for 94.1% of the segments. This

high rate may be due to the ease of aligning session protocols with session transcriptions.

4.2 Correction Stage

In the correction stage, GPT-4o is given the context from subsection 4.1 and the transcription to be corrected, with an extensive, iteratively expanded system prompt specifying usage of the retrieved chunk and additional rules related to peculiarities of the Bernese dialect [2].

The pipeline run with high-compute settings improves the word error rate (WER) from 15.7% to 11.1% when evaluated on 50 minutes of manually transcribed data with temperature set to 0.1 to reduce variability and lower WER (see Figure 5).

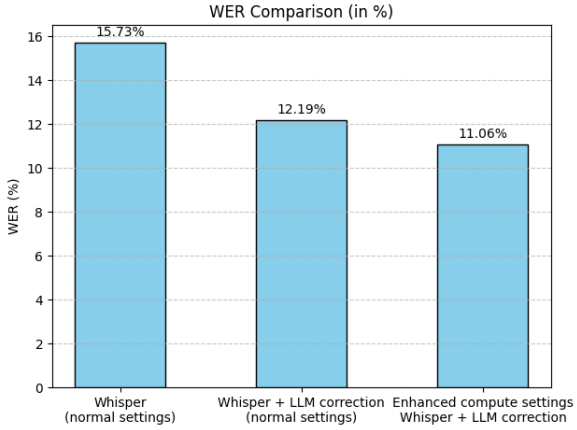

Figure 5: Word Error Rates (WER) for Whisper Large-v3 under three configurations: standard settings, after applying GPT-4o correction, and using high-compute settings (enhanced settings) with GPT-4o correction.

Additionally, when manually inspecting named entities such as places, names, legal references, and political parties, the correctness of named entity transcriptions increases from initial 72.2% with Whisper Large-v3 (52 out of 72) to 100% (72 out of 72) after applying GPT-4o correction.

Table 1 shows an example of the audio, the initial Whisper Large-v3 transcription, the context retrieved, and the output corrected with GPT-4o.

4.3 Evaluation Stage

At this stage, the quality of the transcription is evaluated in the following categories (referred to as judgment tokens hereafter):

[2]Rules include cases such as "vo dr" (audio) to be corrected from "vor der" to "von der" and "mier" (audio) to be corrected from "mir" to "wir".

Table 1: Example audio input, initial transcription with Whisper Large-v3, retrieved context (shortened) given to GPT-4o, and its output. GPT-4o is encouraged to keep the correction as close to the input as possible, so that the data can still be used to train an ASR system that relies on aligned audio and text.

Audio Input (transcribed)
dass ehr au verdaut händ, wenn ehr näbem outo send.
Whisper Large-v3 output (initial transcription)
dass er auch verdauert hat, wenn er neben dem Auto sitzt.
Context retrieved via RAG (given to GPT-4o as help for the correction.)
sodass Sie wieder leicht ernüchtert sind und verdaut haben, wenn Sie beim Auto ankommen werden.
GPT-4o output (final, corrected transcription)
dass Sie auch verdaut haben, wenn Sie neben dem Auto sind.

- **3) Fully correct:** All names, nouns, numbers, and abbreviations are accurately transcribed without any mistakes.
- **2) Minor error** (not affecting key terms)**:** All names, nouns, numbers, and abbreviations are correct. Small grammatical error present (e.g., incorrect conjugation or article).
- **1) Key term error:** At least one name, noun, number, or abbreviation is incorrect in the transcription.
- **0) No relevant excerpts:** The provided excerpt does not contain any relevant content, making evaluation and correction impossible.

Figure 6 presents output of the evaluation stage: 78.0% of transcripts are semantically identical, which means that the context is perfectly reflected in the transcription, after being corrected by GPT-4o.

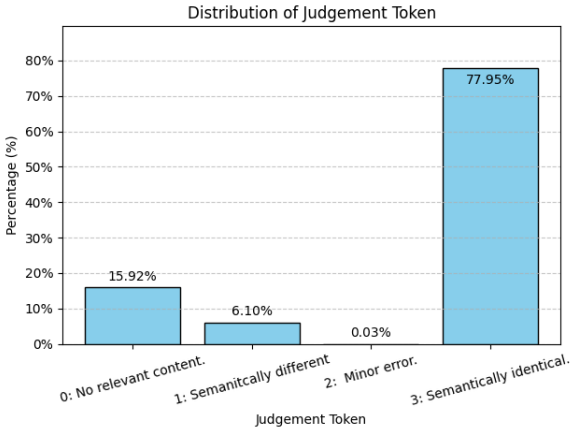

Figure 6: Distribution of the categorization of the final transcription quality using GPT-4o-mini-as-a-judge.

After analyzing 50 minutes of data, we discovered that the judgment category is reliable only when we collapse the label "token 0" into "token 1" and likewise merge "token 2" with "token 3." Grouping the classes this way raises categorization accuracy to 92.2%. Because GPT-4o-mini struggles to decide whether an error is due to missing context or to a genuine semantic change in the transcription, we fuse those tokens for the final data selection.

5 Selecting Data and Train/Test Split

For the construction of the SPC_R high-quality corpus, we combine findings from Section 3.1 (Predicted BLEU) and Section 4.2 (Judgement token) as presented in Figure 7.

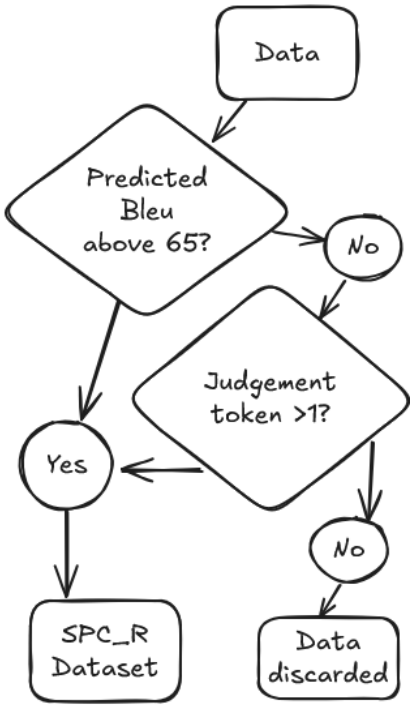

Figure 7: Logic used to build high-quality SPC_R corpus dataset. Size of initial dataset "Data" is 801 hours of audio, size of high-quality dataset "SPC_R" is 751 hours.

We select a Predicted BLEU score threshold of 65 for filtering based on prior research (Cloud) suggesting BLEU score above 60 to be indicative of transcription quality superior to general human levels. By choosing a slightly higher threshold, we reduce the variability indicated by the 95% confidence interval in Figure 2. While this does not guarantee perfect data, (Timmel et al., 2024) shows that imperfect, pseudo-labelled data can improve the quality of ASR models when used in combination with high-quality training data.

This leads to a high quality corpus of 751 hours of Swiss German audio with paired Standard German transcriptions. For the test set, 50 hours are selected with at least a BLEU score of 70 and segments being evaluated as category 3 (as described in Section 4.3). The train/test split is therefore 701/50 hours.

6 Availability and License

The dataset is publicly available on Hugging Face at i4ds/spc_r, the complete codebase (including the prompts) is publicly available on GitHub at i4ds/spc_r.

This dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) License, which allows sharing and adaptation provided that appropriate credit is given and any derivatives are licensed under the same terms.[3]

7 Conclusion

We present SPC_R, transcribed with Whisper Large-v3 on high-compute settings, corrected with context by GPT-4o, and evaluated for quality by GPT-4o-mini. This process results in a corpus of 751 hours of high-quality spoken Swiss German paired with Standard German text.

8 Future Work

There are several promising avenues for further enhancing the Swiss Parliaments Corpus. For instance, incorporating additional data sources beyond the Bernese parliamentary debates could broaden the dialectical and contextual diversity of the dataset, potentially leading to performance and robustness improvements of Swiss German ASR models. Exploring alternative transcription models, especially open source solutions, may offer cost or performance advantages over current approaches based on OpenAI models. Finally, there is also room to work with more nuanced evaluation metrics such as Para_{both} (Paonessa et al., 2023), which better capture semantic fidelity and the accurate transcription of named entities.

9 Limitations

Evaluation Metrics: Our evaluation relies primarily on standard metrics such as BLEU and WER.

[3]For more details, see https://creativecommons.org/licenses/by/4.0/.

These metrics, while useful, do not capture all aspects of transcription quality, as they can be misleading if a sentence conveys the correct semantics using different words, and especially in terms of correctly transcribing named entities, as they don't weight the greater impact of named entity errors on the comprehension of the transcription. In our experience, most of Whisper's errors, which reduce comprehension of the transcription, are now in the named entities, at least in Swiss German.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Google Cloud. Evaluate models | cloud translation. https://cloud.google.com/translate/docs/advanced/automl-evaluate. Accessed: 2025-03-12.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *Preprint*, arXiv:2103.11401.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and pypdf Contributors. 2024. The pypdf library.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jongwook Kim. 2023. Extract confidence. https://github.com/openai/whisper/discussions/1183#discussioncomment-1234567. GitHub Discussion Comment, Accessed: 2025-03-12.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

OpenAI. 2023. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2025-02-28.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Claudio Paonessa, Dominik Frefel, and Manfred Vogel. 2023. Improving metrics for speech translation. *arXiv preprint arXiv:2305.12918*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Christian Scheller, Yanick Schraner, Claudio Paonessa, Larissa Schmidt, Julia Hartmann, Tanja Samardzic, Manfred Vogel, and Mark Cieliebak. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. In preparation.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. In *Proceedings of the Swiss Text Analytics Conference*.

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. PMLR.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Fine-tuning whisper on low-resource languages for real-world applications. *arXiv preprint arXiv:2412.15726*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.