TeluguST-46: A Benchmark Corpus and Comprehensive Evaluation for Telugu-English Speech Translation

Anonymous ACL submission

Abstract

Despite Telugu being spoken by over 80 million people, speech translation research for this morphologically rich language remains severely underexplored. We address this gap by developing a high-quality Telugu–English speech translation benchmark from 46 hours of manually verified CSTD corpus data (30h/8h/8h train/dev/test split). Our systematic comparison of cascaded versus end-to-end architectures shows that while IndicWhisper + IndicMT achieves the highest performance due to extensive Telugu-specific training data, fine-tuned SeamlessM4T models demonstrate remarkable competitiveness despite using significantly less Telugu-specific training data. This finding suggests that with careful hyperparameter tuning and sufficient parallel data (potentially less than 100 hours), end-to-end systems can achieve performance comparable to cascaded approaches in low-resource settings. Our analysis reveals that Wav2Vec-based cascaded systems perform poorly due to limited Telugu ASR training data, while our metric reliability study evaluating BLEU, METEOR, ChrF++, ROUGE-L, TER, and BERTScore against human judgments reveals that traditional metrics provide better quality discrimination than BERTScore for Telugu–English translation. The work delivers three key contributions: a reproducible Telugu–English benchmark, empirical evidence of competitive end-to-end performance potential in low-resource scenarios, and practical guidance for automatic evaluation in morphologically complex language pairs.

1 Introduction

Speech Translation (ST) systems aim to convert spoken input in one language into textual output in another, combining the challenges of automatic speech recognition (ASR) and machine translation (MT). While ST has made significant progress for high-resource language pairs like English–French, low-resource languages from the Indian subcontinent, such as Telugu, remain vastly underexplored.

Telugu is spoken by over 80 million people and is the official language of the Indian states Andhra Pradesh and Telangana. Despite this large speaker base, Telugu–English ST research is hindered by the lack of high-quality speech-text parallel data and robust evaluation protocols. Telugu's agglutinative morphology and rich linguistic structure further complicate the task, often rendering standard evaluation metrics insufficient for capturing translation quality.

Recent efforts such as India's *Bhaashini* initiative and the AI4Bharat consortium have significantly advanced NLP for Indian languages. Models like IndicWhisper (Bhogale et al., 2023), IndicWav2Vec (Javed et al., 2022 (to appear)), and IndicTrans (Gala et al., 2023) have enabled ASR and MT for many Indic languages, including Telugu. These models benefit from extensive Telugu-specific training data, with IndicWhisper trained on over 800 hours of Telugu speech data and IndicMT having substantial exposure to Telugu-English parallel corpora. This language-specific training advantage becomes particularly important in low-resource scenarios.

Contemporary ST systems fall into two broad paradigms. Cascaded systems, which perform ASR followed by MT, benefit from modular design and can leverage language-specific models trained on larger datasets for each component. End-to-end systems, such as SeamlessM4T (Seamless Communication, 2023) and SONAR (Duquenne et al., 2023), directly translate speech to text using unified neural architectures. While end-to-end approaches often show advantages for high-resource language pairs due to joint optimization, their effectiveness on low-resource Indic language pairs may be limited by insufficient language-specific training data. For instance, SeamlessM4T's Telugu-English training data comprises only 466 hours compared to the

800+ hours available to IndicWhisper for Telugu ASR alone.

Equally important is the question of evaluation. While BLEU (Papineni et al., 2002) remains a widely used metric, it is known to underperform on morphologically rich languages. Alternatives such as ChrF++ (Popović, 2015), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) offer different perspectives on translation quality. TER (Snover et al., 2006) provides another angle on edit distance. However, their reliability and discriminative power in Telugu–English ST scenarios has not been systematically studied, particularly regarding which metrics can effectively distinguish between different model performances. **This paper addresses these gaps through three major contributions:** (1) We release 46 hours of manually verified Telugu–English parallel data curated from the CSTD corpus (Mirishkar et al., 2021), facilitating reproducible benchmarks for ST in a low-resource Indian (2) We evaluate the performance of both cascaded (IndicWhisper+IndicMT, IndicWav2Vec+IndicMT) and end-to-end models (SeamlessM4T, SONAR), providing empirical evidence that while cascaded approaches with extensive language-specific training data achieve superior performance, carefully fine-tuned end-to-end systems show promising competitiveness with limited data. (3) We analyze the reliability and discriminative power of six standard automatic metrics (BLEU, METEOR, ChrF++, ROUGE, TER, and BERTScore) by comparing them with ChatGPT-style human evaluations.

We also investigate the effect of mixing verified and unverified data during fine-tuning, offering insights into corpus construction strategies under resource constraints. Together, these efforts provide a solid foundation for Telugu–English ST benchmarking and offer broader implications for other morphologically rich, underrepresented languages. All resources—data splits, evaluation scripts, and system outputs—are made publicly available to foster reproducible and inclusive research in multilingual speech translation.

2 Dataset Creation

We construct a high-quality Telugu-English speech translation dataset from the CSTD corpus (Mirishkar et al., 2021), addressing cross-lingual alignment and dialectal variation challenges through systematic quality control.

2.1 Source Dataset Selection

We utilize the CSTD corpus, comprising 2,000 hours of transcribed Telugu audio across three dialectal variants: Telangana, Rayalaseema, and Andhra. We strategically selected a 50-hour subset maintaining dialectal diversity while ensuring computational feasibility for manual verification processes.

2.2 Translation Pipeline and Quality Control

2.2.1 Initial Translation Generation

Telugu transcripts were extracted from the 50-hour subset and translated to English using IndicTrans MT, selected for its effectiveness with morphologically rich Indian languages and Telugu-English parallel training data.

2.2.2 Manual Verification Protocol

Each audio-translation pair underwent evaluation by native Telugu speakers using a 5-point Likert scale:

- Score 5: Perfect semantic alignment with natural English expression

- Score 4: Minor variations, core meaning preserved

- Score 3: Acceptable translation with moderate modifications required

- Score 2: Significant gaps requiring substantial revision

- Score 1: Fundamental misalignment requiring complete retranslation

Pairs scoring below 3 received manual correction. Audio segments with multiple speakers, background noise, or recording artifacts were eliminated through automatic detection and manual inspection.

2.2.3 Dataset Composition

Following quality control, we obtained 46 hours of verified parallel data (92% retention rate), partitioned as Training Set: 30 hours (65.2%),Development Set: 8 hours (17.4%), Test Set: 8 hours (17.4%)

Metric	Description
BLEU	Focuses on lexical matching and word order.
ChrF++	Captures sub-word similarities and handles morphological variations.
BERTScore	Captures semantic meaning beyond surface form matching.
TER	Measures minimum number of edits (insertions, deletions, substitutions, shifts) required to transform hypothesis into reference.
ROUGE-L	Captures structural similarity and word order.
METEOR	Incorporates linguistic knowledge and semantic equivalence.

Table 1: Overview of evaluation metrics and their calculation approaches for assessing translation quality.

2.2.4 Key Challenges

Primary challenges included transcription inconsistencies from dialectal variations, translation ambiguities from Telugu's morphological richness, and temporal alignment drift. These were addressed through standardization protocols, manual correction procedures, and systematic alignment verification.

3 Speech Translation Architectures and Training Setup

3.1 Model Architectures

We evaluate four speech translation architectures spanning two major paradigms: cascaded and end-to-end systems.

The cascaded models follow a modular design and benefit from component-specific training advantages. **IndicWhisper + IndicMT** (cascad-1) integrates IndicWhisper, an ASR model fine-tuned on over 800 hours of Telugu speech data, with IndicTrans, a neural machine translation system with extensive Telugu-English parallel training. This combination leverages substantial language-specific training data for both ASR and MT components. **IndicWav2Vec + IndicMT** (cascad-2) combines a Telugu-specific Wav2Vec-based speech encoder with IndicTrans. However, this configuration faces limitations as the Wav2Vec model [1] has been trained on significantly limited Telugu data compared to IndicWhisper, potentially leading to incorrect transcriptions that negatively impact downstream translation quality.

The end-to-end models include **SeamlessM4T** and **SONAR**. SeamlessM4T is a massively multilingual model that directly maps speech to text using a unified architecture, but with limited Telugu-English training data (466 hours) compared to

the cascaded components. Despite this limitation, our experiments reveal that SeamlessM4T shows promising potential when fine-tuned with careful hyperparameter selection. SONAR is a unified multilingual encoder that produces sentence-level representations for over 200 languages across both speech and text modalities, though it similarly lacks the extensive Telugu-specific training of the cascaded components.

3.2 Training Configurations

We investigate the impact of training data quality by fine-tuning SeamlessM4T under three distinct configurations. The *Verified Only* setting uses 30 hours of manually verified data as the baseline. The *Balanced* configuration expands the dataset to 45 hours by combining the 30 hours of verified data with an additional 15 hours of unverified speech-text pairs. Finally, the *Unbalanced* configuration further increases the total to 73 hours by incorporating a larger proportion of unverified data, allowing us to examine the effects of noisy data on model performance.

For hyperparameter optimization, we explore learning rates of 1×10^{-5} and 1×10^{-6}, batch sizes of 10, 32, and 64, and warmup steps set to either 100. These variations are used to identify stable training regimes that balance learning efficiency and generalization performance across the data configurations.

4 Evaluation Metrics

We employ six automatic metrics capturing different translation quality aspects. Table 1 provides an overview of these metrics and their computational approaches, highlighting how each metric addresses different dimensions of translation quality assessment.

---

[1]https://huggingface.co/anuragshas/wav2vec2-large-xlsr-53-telugu

Model	BLEU	ChrF++	TER (↓)	ROUGE-L	METEOR	BERTScore	ChatGPT (%)
cascad-1	**23.99**	**47.69**	**82.00**	**48.47**	**47.19**	91.9	**58.3**
cascad-2	8.3	28.27	99.8	28.23	20.17	89.12	41.2
Configuration 0	13.9	38.9	89.9	38.7	39.5	89.5	48.1
configuration 1	14.1	39.8	89.5	40.2	40.2	89.1	49.2
configuration 2	16.9	42.87	86.0	42.6	43.01	90.1	51.0
configuration 3	14.9	40.1	87.7	41.2	41.6	89.7	50.4
configuration 4	17.1	42.9	85.9	42.7	43.8	90.9	51.3
configuration 5	15.8	41.8	86.0	41.6	42.4	89.9	50.9
SONAR	10.9	35.8	98.7	36.4	35.6	88.3	49.2

Table 2: Automatic metric scores across cascaded and end-to-end models. Cascaded models, particularly IndicWhisper + IndicMT, significantly outperform end-to-end approaches across most metrics.

4.1 BLEU Score

Measures precision-based n-gram overlap with brevity penalty, emphasizing lexical similarity.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

where p_n is n-gram precision and BP is brevity penalty.

4.2 ChrF++

Evaluates character-level F-score, effective for morphologically rich languages like Telugu.

$$\text{ChrF++} = \frac{(1 + \beta^2) \cdot chrP \cdot chrR}{\beta^2 \cdot chrP + chrR} \quad (2)$$

where $chrP$ and $chrR$ are character-level precision and recall.

4.3 BERTScore

Computes semantic similarity using contextual embeddings, capturing meaning beyond surface form.

$$\text{BERTScore} = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} \mathbf{x}_i^\top \mathbf{y}_j \quad (3)$$

where $\mathbf{x_i}$ and $\mathbf{y_j}$ are BERT embeddings of tokens.

4.4 Translation Edit Rate (TER)

Measures minimum edit operations required for transformation (lower is better).

$$\text{TER} = \frac{\text{Number of Edits}}{\text{Average Reference Length}} \quad (4)$$

4.5 ROUGE

Computes recall-oriented n-gram overlap. We use ROUGE-1, ROUGE-2, and ROUGE-L variants.

$$\text{ROUGE-N} = \frac{\text{Number of matching n-grams}}{\text{Total n-grams in the reference}} \quad (5)$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \quad (6)$$

where P is precision, R is recall, and $\beta = 1$.

4.6 METEOR

Combines precision and recall using exact, stem, synonym, and paraphrase matching with linguistic equivalence.

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \cdot (1 - \text{Penalty}) \quad (7)$$

where P and R are weighted precision and recall, and $Penalty$ accounts for word order differences.

5 Experimental Setup

5.1 Baseline Model Evaluation

Our experimental framework comprised two distinct categories of speech translation models: cascaded and end-to-end architectures. For the cascaded approaches, we evaluated two configurations: IndicWhisper + IndicMT and IndicWav2Vec2.0 + IndicMT. Additionally, we assessed **SONAR** and **Seamless** as our end-to-end baseline models. Initial performance evaluation was conducted by directly inferencing these models on our designated test dataset without any domain-specific adaptations. This provided baseline performance metrics that informed subsequent fine-tuning strategies.

5.2 Fine-Tuning Methodology

Based on the initial inference results, which showed end-to-end models underperforming cascaded approaches, we implemented a systematic fine-tuning approach (Akkiraju et al., 2025) using the SeamlessM4T model to investigate whether targeted fine-tuning could bridge this performance gap . The fine-tuning process was designed to investigate the impact of training data composition and hyperparameter configurations on model performance.

5.3 Training Data Configurations

We conducted fine-tuning experiments across multiple data configurations to assess the relationship between data quality, quantity, and model performance:

- **Configuration 0** : SeamlessM4T large model with Direct Inference

- **Configuration 1: Verified Data Only**
  Utilized exclusively verified, high-quality training data to establish performance under clean training conditions.

- **Configuration 2: Mixed Data (15 hours)**
  Combined verified data with 15 hours of unverified data to evaluate the impact of limited data augmentation using noisy samples.

- **Configurations 3–5: Extended Mixed Data (80 hours)**
  Employed a larger dataset combining verified and unverified data (80 hours total) to investigate the effects of hyperparameter variations at scale.

  Configuration 3: Learning rate = 1×10^{-5}, batch size = 10, representing an aggressive learning setup with smaller batch processing.

  Configuration 4: Learning rate = 1×10^{-6}, batch size = 10, combining conservative learning with smaller batch sizes to balance stability and computational efficiency.

  Configuration 5: Learning rate = 1×10^{-6}, batch size = 32, implementing a conservative learning strategy with larger batch size for improved gradient stability.

6 Results

We evaluate cascaded and end-to-end speech translation models using six automatic metrics—BLEU, ChrF++, TER, ROUGE-L, METEOR, and BERTScore—alongside ChatGPT-style human scoring to assess model performance and metric reliability for Telugu–English speech translation. Table 2 presents results across all configurations.

Our findings highlight the performance dynamics across architectural choices. **IndicWhisper + IndicMT** achieves the best overall performance with 23.99 BLEU, 47.69 ChrF++, and 47.19 METEOR, clearly benefiting from over 800 hours of Telugu ASR training and substantial Telugu-English parallel corpora in the MT stage. This cascaded setup demonstrates the importance of component-specific language exposure in low-resource scenarios.

Interestingly, **fine-tuned SeamlessM4T** models perform competitively despite limited Telugu-specific supervision. **Configuration 4** achieves 17.1 BLEU and 43.82 METEOR—only a 29% BLEU drop from the cascaded baseline—while using under 30 hours of verified training data and 43 hours of unverified training data. This result suggests that with effective fine-tuning, end-to-end models can approach cascaded system performance, offering a scalable alternative where monolithic training pipelines are preferred.

By contrast, **Wav2Vec + IndicMT** underperforms across all metrics (8.3 BLEU, 28.27 ChrF++, 20.17 METEOR), largely due to inadequate Telugu ASR training data. This discrepancy underscores the vital role of ASR quality in cascaded speech translation pipelines. Additionally, **SONAR** achieves only moderate performance (10.9 BLEU, 35.6 METEOR), highlighting that not all end-to-end multilingual systems generalize well to morphologically rich Indian languages. Among end-to-end SeamlessM4T variants, **Configurations 2 and 4** deliver the best results. Configuration 2 achieves 16.9 BLEU and 43.01 METEOR, closely trailing Configuration 4. These results reflect the effectiveness of balanced data curation and learning rate tuning in maximizing translation quality with limited resources.

6.1 Metric Reliability Analysis

To evaluate which metrics align best with human perception, we compute the Pearson correlation coefficient (PCC) between ChatGPT scores and each automatic metric.

- **ROUGE-L** ($r = 0.96$) and **ChrF++** ($r =$

Metric	PCC with ChatGPT
BLEU	0.9417
ChrF++	0.9479
TER	-0.8197
ROUGE-L	0.9645
METEOR	0.9022
BERTScore	0.7372

Table 3: Pearson correlation (PCC) between ChatGPT-style scores and automatic metrics across model configurations.

0.95) show the strongest agreement with human judgment, making them suitable for low-resource ST evaluation.

- **BLEU** ($r = 0.94$) and **METEOR** ($r = 0.90$) also show strong correlations, reinforcing their applicability in this domain.

- **BERTScore** ($r = 0.73$) shows lower alignment across configurations, suggesting limited discriminative power for morphologically rich translations.

- **TER** ($r = -0.82$) exhibits a strong inverse correlation, as expected from an error-based metric, further validating its interpretability in this context.

7 Conclusion

This work presents a comprehensive benchmark for Telugu–English speech translation, releasing 46 hours of verified data from the CSTD corpus. Our experiments reveal important insights into the trade-offs between cascaded and end-to-end architectures in low-resource settings. While cascaded systems like IndicWhisper + IndicMT achieve the best results (23.99 BLEU, 47.19 METEOR), fine-tuned end-to-end models such as SeamlessM4T demonstrate notable competitiveness (17.1 BLEU, 43.82 METEOR) despite relying on significantly less Telugu-specific training data.

This 29% BLEU gap underscores the strong inductive bias provided by Telugu-specific training in cascaded models. However, it also highlights the scalability potential of end-to-end systems: with less than 100 hours of parallel fine-tuning data, SeamlessM4T models can approach the performance of cascaded systems. The weak performance of Wav2Vec + IndicMT (8.3 BLEU,

28.27 ChrF++) further emphasizes the necessity of sufficient ASR exposure for effective cascaded ST.

Our metric reliability analysis demonstrates that **ROUGE-L** and **ChrF++** correlate most strongly with human judgments (Pearson $r = 0.96$ and $r = 0.95$, respectively), suggesting their effectiveness in evaluating low-resource, morphologically rich translations. **BLEU** and **METEOR** also show strong alignment ($r > 0.90$), reinforcing their continued relevance in this setting. In contrast, **BERTScore** exhibits lower correlation ($r = 0.73$), indicating limited sensitivity to performance variation across configurations. As expected, **TER** shows strong inverse correlation ($r = -0.82$), aligning well with human judgments when interpreted as an error-based metric.

Although cascaded systems benefit from large language-specific resources (e.g., 800+ hours for IndicWhisper), our findings indicate that carefully fine-tuned end-to-end systems, with modest high-quality data, can deliver competitive performance. Future work should explore monolingual augmentation, domain-specific adaptation (e.g., medical ST), and multilingual pretraining across related Indian languages to further enhance end-to-end ST in underrepresented contexts.

Overall, this benchmark lays the foundation for reproducible research in Indic speech translation, encourages deeper evaluation methodology, and demonstrates that with thoughtful design, end-to-end systems offer a viable path forward in low-resource speech translation scenarios.

8 Limitations and Future Work

8.1 Limitations

Our study presents few limitations that should be considered when interpreting the results. The primary limitation is the constrained dataset size of 46 hours, which, while manually verified for quality, limits our ability to fully understand the relationship between data quantity and model performance. This constraint prevents us from conducting comprehensive scaling experiments to determine optimal data requirements for different model architectures. The potential for including more verified data from the CSTD corpus remains unexplored, which could provide valuable insights into how data size impacts translation quality across cascaded and end-to-end approaches.Another significant limitation is our focus on a single end-to-end model (SeamlessM4T) for fine-tuning experiments.

6

While this provided insights into the potential of end-to-end approaches, exploring other architectures could yield different conclusions about the scalability and effectiveness of unified speech translation systems for Telugu.

8.2 Future Work

Several promising research directions emerge from this work that could significantly advance Telugu speech translation and broader low-resource ST research.

Dataset Expansion and Quality Analysis: A critical next step involves scaling the verified dataset to investigate the data size-performance relationship systematically. Expanding from the current 46 hours to 100+ hours of verified Telugu-English parallel data would enable more robust conclusions about optimal training data requirements for different model architectures.

Advanced Fine-tuning Strategies for End-to-End Models: Given the promising results of SeamlessM4T fine-tuning, future research should explore more sophisticated adaptation techniques including domain adaptation strategies, few-shot learning approaches, and parameter-efficient fine-tuning methods such as LoRA or adapters.

Cross-lingual Transfer and Multilingual Extensions: Investigating how insights from Telugu-English speech translation can transfer to other Dravidian languages (Tamil, Kannada, Malayalam) could provide valuable guidance for developing ST systems across the Indian subcontinent. This includes exploring multilingual training strategies and cross-lingual transfer learning approaches that leverage shared linguistic properties among related languages.

References

Bhavana Akkiraju, Aishwarya Pothula, Santosh Kesiraju, and Anil Vuppala. 2025. IIITH-BUT system for IWSLT 2025 low-resource Bhojpuri to Hindi speech translation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 333–339, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Kaushal Santosh Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M. Khapra, and Pratyush Kumar. 2023. Vistaar: Diverse benchmarks and training sets for indian language asr. *Preprint*, arXiv:2305.15386.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022 (to appear). Towards building asr systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ganesh S Mirishkar, Meher Dinesh Naroju, Sudhamay Maity, Prakash Yalla, Anil Kumar Vuppala, and 1 others. 2021. Cstd-telugu corpus: Crowd-sourced approach for large-scale speech data collection. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 511–517. IEEE.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Yu-An Chung Mariano Coria Meglioli David Dale Ning Dong Mark Duppenthaler Paul-Ambroise Duquenne Brian Ellis Hady Elsahar Justin Haaheim John Hoffman Min-Jae Hwang Hirofumi Inaguma Christopher Klaiber Ilia Kulikov Pengwei Li Daniel Licht Jean Maillard Ruslan Mavlyutov Alice Rakotoarison Kaushik Ram Sadagopan Abinesh Ramakrishnan Tuan Tran Guillaume Wenzek Yilin Yang

7

Ethan Ye Ivan Evtimov Pierre Fernandez Cynthia Gao Prangthip Hansanti Elahe Kalbassi Amanda Kallet Artyom Kozhevnikov Gabriel Mejia Robin San Roman Christophe Touret Corinne Wong Carleigh Wood Bokai Yu Pierre Andrews Can Balioglu Peng-Jen Chen Marta R. Costa-jussà Maha Elbayad Hongyu Gong Francisco Guzmán Kevin Heffernan Somya Jain Justine Kao Ann Lee Xutai Ma Alex Mourachko Benjamin Peloquin Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Anna Sun Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Mary Williamson Seamless Communication, Loïc Barrault. 2023. Seamless: Multilingual expressive and streaming speech translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

8