Fine-tuned NLLB-200 for Bidirectional Translation between Mexican Sign Language Glosses and Spanish: FisBioUNAM at MSLG-SPA 2026

David Alexis García-Espinosa[1,*], Luis Eduardo Flores-Luna[1] and Ángel Andrés Moreno-Sánchez[1]

[1]*Universidad Nacional Autónoma de México, Facultad de Ciencias, Ciudad de México, MX*

Abstract

This paper describes the FisBioUNAM system for the MSLG-SPA 2026 shared task at IberLEF 2026 [1], targeting bidirectional translation between Mexican Sign Language (LSM) glosses and Spanish. We fine-tune NLLB-200-distilled-600M on 490 aligned gloss–Spanish pairs augmented to 968 via LLM-based synthetic generation. Model selection was empirical: NLLB-200 outperformed mT5-small by $4\times$ on BLEU (25.74 vs. 6.60) in internal benchmarks. In the official evaluation, we ranked 9th of 20 teams on Subtask A (gloss-to-Spanish, Task Score 0.3724) and 9th of 19 on Subtask B (Spanish-to-gloss, Task Score 0.1003). We analyze the pipeline decisions, the impact of data augmentation in this low-resource setting, and a characteristic failure mode where the model generates Spanish instead of glosses in Subtask B.

Keywords

Mexican Sign Language, gloss translation, NLLB-200, low-resource machine translation, data augmentation, IberLEF 2026

1. Introduction

Mexican Sign Language (Lengua de Señas Mexicana, LSM) serves an estimated 300,000 to 700,000 signers in Mexico but remains underrepresented in NLP research, primarily due to the scarcity of parallel corpora aligning sign language representations with written text.

Gloss notation provides a practical intermediate representation for sign languages, transcribing each sign as an upper-case token roughly corresponding to its meaning. While glosses sacrifice the rich spatial and non-manual components of signing, they enable text-based sequence-to-sequence modeling and serve as a bridge between video-based sign recognition and written language generation [2].

The MSLG-SPA 2026 shared task [3], hosted at IberLEF 2026 [1], is the first competitive evaluation for bidirectional translation between LSM glosses and Spanish, comprising two subtasks:

- **Subtask A (MSLG2SPA):** Translate LSM gloss sequences into fluent Spanish sentences.
- **Subtask B (SPA2MSLG):** Translate Spanish sentences into LSM gloss sequences.

We fine-tune NLLB-200-distilled-600M [4] for both directions, selected after comparing it against mT5-small [5] on internal benchmarks. We augmented the training set from 490 to 968 pairs using synthetic generation via Claude [6]. The system placed 9th on both subtasks. We report the pipeline decisions, what worked, and what did not—including a failure mode specific to the gloss generation direction.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

davale@ciencias.unam.mx (D. A. García-Espinosa)

0000-0001-6141-407X (D. A. García-Espinosa); 0009-0007-6087-8658 (L. E. Flores-Luna); 0009-0007-8978-6087 (Á. A. Moreno-Sánchez)

Table 1
LSM gloss notation conventions used in the MSLG-SPA 2026 dataset.

Convention	Example	Meaning
dm-NAME	dm-MARÍA	Fingerspelling (proper noun)
+MUJER	HERMANO+MUJER	Feminization (e.g., "sister")
#ACRONYM	#TV, #SEP	Abbreviation / acronym
HYPHEN-COMP	LICENCIA-DE-CONDUCIR	Multi-word concept
DUAL	NOSOTROS-DE-DOS	Dual pronoun
suffix	HOSPITALacá	Positional classifier
REPETITION	EJERCER EJERCER	Intensification

2. Related Work

Neural sign language translation. Camgöz et al. [2] introduced the PHOENIX-2014T benchmark for German Sign Language (DGS) and proposed an encoder–decoder architecture for end-to-end sign language translation from video to text. Subsequent work by Yin and Read [7] demonstrated improvements through pre-training strategies and better tokenization. De Coster et al. [8] showed that frozen pretrained transformer encoders can be effectively adapted for sign language translation, motivating the use of large pretrained multilingual models.

Multilingual pretrained models. The NLLB (No Language Left Behind) project [4] released a family of sequence-to-sequence models covering 200 languages, with the distilled 600M-parameter variant providing strong translation quality at moderate computational cost. While LSM glosses are not among NLLB's training languages, the model's extensive Spanish capabilities and multilingual transfer learning make it a promising foundation for fine-tuning on low-resource translation pairs. The mT5 model [5], a multilingual variant of T5 [9], provides an alternative pretrained encoder–decoder architecture but proved less effective in our experiments.

Data augmentation for low-resource MT. Data augmentation has been widely explored for low-resource machine translation, including back-translation [10], paraphrase generation, and synthetic data creation. In our setting, we employ a large language model to generate additional gloss–Spanish pairs following LSM conventions, an approach that has shown promise in other low-resource scenarios.

Evaluation metrics. We rely on the standard metrics adopted by the shared task: BLEU [11], METEOR [12], chrF [13], and COMET [14]. COMET is only applicable to Subtask A, as gloss sequences are not natural language sentences suitable for neural quality estimation.

3. Task and Data

3.1. Task Description

The MSLG-SPA 2026 shared task [3] provides a parallel corpus of LSM gloss sequences and Spanish sentences. Systems must perform translation in both directions. Evaluation uses z-score normalized metrics, with the Task Score for each subtask defined as the mean of its constituent z-scores. Subtask A uses four metrics (BLEU, METEOR, chrF, COMET), while Subtask B uses three (BLEU, METEOR, chrF).

3.2. LSM Gloss Conventions

LSM glosses follow specific conventions that distinguish them from standard written text. Table 1 summarizes the main notational devices.

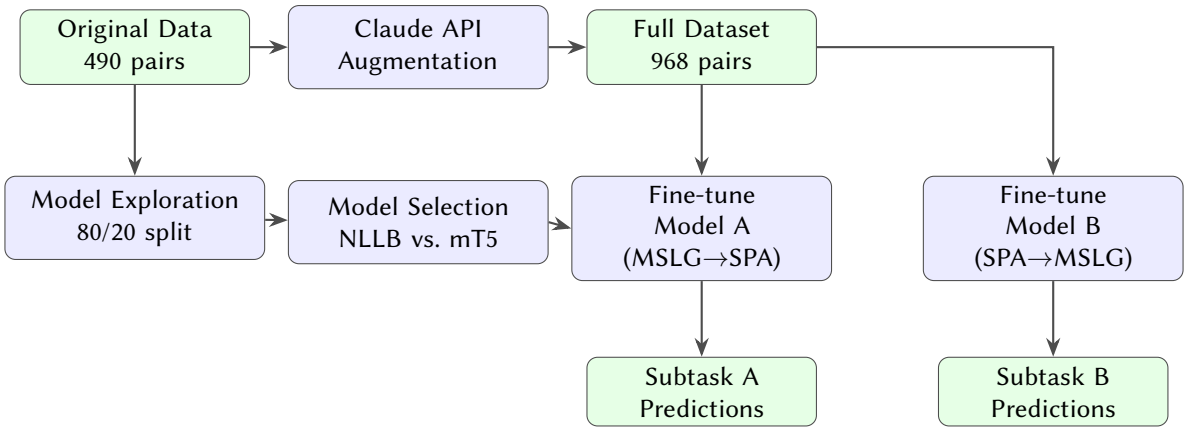

Figure 1: Overview of the FisBioUNAM system pipeline. Green nodes represent data artifacts; blue nodes represent processing stages.

These conventions pose significant challenges for neural models, which must learn to generate structured tokens (e.g., `dm-` prefixes for proper nouns) that rarely or never appear in their pretraining data.

3.3. Dataset Statistics

The organizers provided 490 aligned MSLG–SPA training pairs. For evaluation, two disjoint test sets were released: 245 gloss sequences for Subtask A and 245 Spanish sentences for Subtask B. No dedicated development set was provided. The training pairs exhibit diverse sentence structures, ranging from simple declarative statements to complex multi-clause sentences.

4. System Description

Figure 1 provides an overview of our system pipeline, from data augmentation through model selection, training, and inference.

4.1. Data Augmentation

Given the extremely small training set (490 pairs), we employed Anthropic's Claude API [6] to generate approximately 478 additional synthetic MSLG–SPA pairs. The generation prompt provided the model with:

1. A description of LSM gloss conventions (Table 1).
2. A sample of existing training pairs as few-shot examples.
3. Instructions to produce diverse, grammatically correct pairs covering varied topics and sentence structures.

This augmentation nearly doubled our training set to 968 pairs. The synthetic pairs were manually spot-checked for adherence to gloss conventions but were not exhaustively validated due to time constraints. For the final submission, we trained on all 968 pairs without holding out a validation set, in order to maximize training data in this extremely low-resource setting.

4.2. Model Selection

Before committing to a final architecture, we conducted a systematic exploration comparing two multilingual pretrained models on an 80/20 split of the original 490 training pairs, evaluating on the MSLG→SPA direction. Table 2 and Figure 2 present the results.

Table 2
Model exploration on 80/20 split of 490 pairs (MSLG→SPA). NLLB-200-600M outperforms mT5-small by 4× on BLEU.

Model	BLEU	METEOR	chrF
mT5-small (30 ep, lr 10^{-4})	6.60	28.58	27.01
mT5-small (50 ep, lr 3×10^{-5})	4.49	19.53	20.02
NLLB-200-600M (30 ep, lr 3×10^{-5})	**25.74**	**56.83**	**56.47**

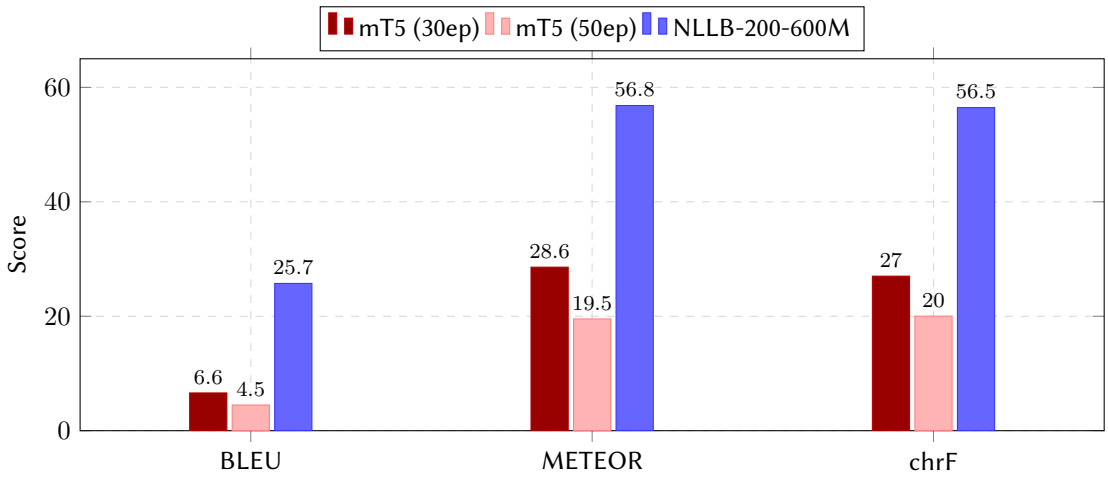

Figure 2: Model exploration: NLLB-200-600M vs. mT5-small on the MSLG→SPA validation split.

NLLB-200 achieved BLEU 25.74 vs. mT5's best of 6.60, with proportional gains on METEOR and chrF. The gap reflects NLLB's translation-specific pretraining versus mT5's general-purpose text-to-text design. We selected NLLB-200-600M based on these results.

4.3. Model Architecture and Training

NLLB-200-distilled-600M [4] is a sequence-to-sequence Transformer [15] model distilled from the larger NLLB-200-3.3B. It supports 200 languages using language-specific tokens prepended to the input. Since LSM glosses are not a supported language, we used `spa_Latn` (Spanish, Latin script) as both the source and target language token for both directions. This leverages the model's strong Spanish capabilities while allowing the encoder to adapt to gloss-formatted input during fine-tuning.

We trained two separate models using the HuggingFace Transformers library [16]:

- **Model A (MSLG→SPA):** Source = gloss sequences, Target = Spanish sentences.
- **Model B (SPA→MSLG):** Source = Spanish sentences, Target = gloss sequences.

Table 3 lists the hyperparameters used for both models.

The choice of Adafactor [17] as the optimizer was motivated by its memory efficiency, which is critical when fine-tuning a 600M-parameter model on a consumer-grade GPU. A warmup ratio of 0.1 with 30 epochs provides approximately 3 warmup epochs, allowing the model to gradually adapt to the new domain without catastrophic forgetting.

5. Results

5.1. Official Results: Subtask A (MSLG→SPA)

Table 4 presents the official leaderboard for Subtask A. FisBioUNAM ranked 9th out of 20 participating teams, with a Task Score of 0.3724. Our absolute metric scores were BLEU 31.28, METEOR 0.5171, chrF 62.09, and COMET 0.8103.

Table 3
Hyperparameters for fine-tuning NLLB-200-distilled-600M.

Parameter	Value
Base model	facebook/nllb-200-distilled-600M
Epochs	30
Learning rate	3×10^{-5}
Batch size	4
Gradient accumulation steps	2 (effective batch = 8)
Max sequence length	64 tokens
Beam search	4 beams
Optimizer	Adafactor
Warmup ratio	0.1
Weight decay	0.01
Mixed precision	FP16 (on CUDA)
Hardware	Google Colab T4 (15.6 GB VRAM)
Source/target lang	`spa_Latn` / `spa_Latn`

Table 4
Official Subtask A (MSLG→SPA) results. Top 10 teams shown. FisBioUNAM is highlighted.

Rank	Team	Task Score	BLEU	METEOR	chrF	COMET
1	CICESE-LCDAA	1.3421	54.03	0.7191	76.56	0.8971
2	AxoloTux	1.3400	52.70	0.7272	76.40	0.9020
3	UAM_ChineseRoom	1.1890	50.53	0.6808	73.93	0.8903
4	UCOUCPlenitas	1.0628	47.62	0.6515	71.86	0.8823
5	VerbaNexAI	1.0339	44.96	0.6488	72.35	0.8862
6	Ludwig	1.0055	45.49	0.6473	70.66	0.8804
7	ComunicadorInnato	0.8811	38.70	0.6416	70.40	0.8733
8	LSM-INAOE	0.4105	32.11	0.5290	60.54	0.8280
9	**FisBioUNAM**	**0.3724**	**31.28**	**0.5171**	**62.09**	**0.8103**
10	RETUYT-INCO	0.2043	28.51	0.4982	58.70	0.7805

Table 5
Official Subtask B (SPA→MSLG) results. Top 10 teams shown. FisBioUNAM is highlighted.

Rank	Team	Task Score	BLEU	METEOR	chrF
1	AxoloTux	1.7577	26.49	0.5974	68.29
2	VerbaNexAI	1.4466	23.79	0.5665	65.39
3	UAM_ChineseRoom	1.2724	21.36	0.5612	64.21
4	CICESE-LCDAA	1.0072	21.19	0.5104	60.37
5	RETUYT-INCO	0.5705	17.77	0.4562	56.71
6	ReCore	0.4218	17.46	0.4253	55.21
7	UCOUCPlenitas	0.2329	10.82	0.4518	57.80
8	LSM-INAOE	0.1125	13.88	0.3796	55.39
9	**FisBioUNAM**	**0.1003**	**14.51**	**0.4209**	**49.73**
10	DunderMifflin	0.0405	14.47	0.3904	50.76

5.2. Official Results: Subtask B (SPA→MSLG)

Table 5 presents the official leaderboard for Subtask B. FisBioUNAM again ranked 9th, this time out of 19 teams, with a Task Score of 0.1003.

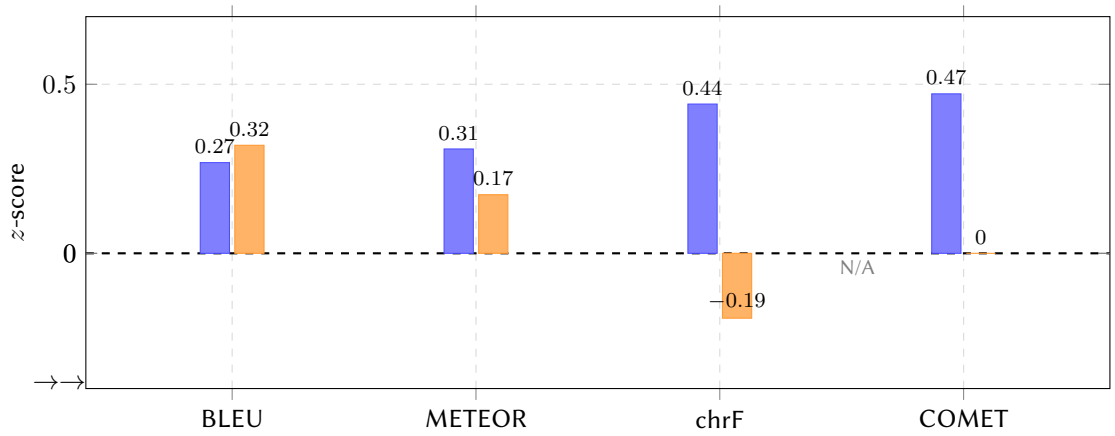

Figure 3: FisBioUNAM per-metric z-scores for both subtasks. Subtask B has no COMET score (glosses are not natural language). The negative chrF z-score for Subtask B reveals difficulty in character-level gloss formatting.

Table 6
Sample translations produced by FisBioUNAM.

Input	Output
Subtask A: MSLG→SPA	
ELLA TENER BUSTO PEQUEÑO	Ella tiene un busto pequeño. ✓
MI HERMANO #TV ÉL REPORTERO	Mi hermano es reportero de televisión. ✓
AÑO PASADO BOTSUANA YO VIAJAR	El año pasado viajé a Botswana. ✓
dm-JUAN SUYO PERRO YA ABANDONAR	Juan abandonó a su perro. ✓
Subtask B: SPA→MSLG	
El gobernador apoyó la construcción de una escuela para alumnos sordos.	GOBERNADOR APOYO CONSTRUCCIÓN ESCUELA PARA ALUMNOS SORDOS ✓
Javier es mi concuño.	dm-JAVIER YO CONCUÑO ✓
Yo compré una camisa elegante.	YO YA COMPRAR CAMISA ELEGANTE ✓
Esa marca es excelente.	Esa marca es excelente. ✗

5.3. Metric Analysis

Figure 3 visualizes our per-metric z-scores across both subtasks. For Subtask A, performance is relatively balanced across all four metrics, with chrF and COMET being our strongest indicators ($z = 0.44$ and $z = 0.47$ respectively). For Subtask B, there is a notable disparity: BLEU is our strongest metric ($z = 0.32$), while chrF is actually negative ($z = -0.19$), indicating below-average character-level overlap, likely due to the model's difficulty in consistently producing properly formatted gloss tokens.

6. Analysis and Discussion

6.1. Qualitative Analysis of Translations

Table 6 presents sample translations from both subtasks, illustrating both successes and failures.

For Subtask A, the model handles short declarative sentences well, correctly resolving gloss conventions such as `dm-` (fingerspelling) and `#` (acronyms), and producing fluent Spanish with appropriate articles, conjugations, and word order. For Subtask B, while many outputs correctly adopt gloss conventions (upper case, omitted articles, SOV-like order), a critical failure mode emerges: the model occasionally produces Spanish text verbatim instead of converting to glosses, as seen in the last example of Table 6.

6.2. Asymmetry Between Subtasks

Our Task Score is $3.7\times$ higher for Subtask A (0.3724) than Subtask B (0.1003). In the gloss-to-Spanish direction, the model leverages its pretrained Spanish generation capabilities. In the reverse direction, it must produce a structured notation absent from pretraining, relying entirely on 968 fine-tuning examples. This asymmetry is consistent across the leaderboard: all teams score higher on Subtask A.

6.3. Impact of Data Augmentation

Our augmentation strategy nearly doubled the training data from 490 to 968 pairs. While we cannot directly measure the impact of augmentation on our final submission (since we trained on all 968 pairs without ablation), the exploration results on 490 pairs yielded a BLEU of 25.74 (80/20 split), while the final model on 968 pairs achieved a BLEU of 31.28 on the official test set. Although these numbers are not directly comparable (different test sets, different data sizes), the trend is consistent with the augmentation contributing positively to performance.

A limitation of our augmentation approach is that synthetic pairs generated by a large language model may not perfectly replicate the distribution of naturally occurring LSM glosses. Some gloss conventions (e.g., positional classifiers, dual pronouns) may be underrepresented in the augmented data if the language model did not fully internalize these patterns.

6.4. Comparison with Top Systems

The top team on Subtask A (CICESE-LCDAA) achieved BLEU 54.03 vs. our 31.28; on Subtask B (AxoloTux), 26.49 vs. our 14.51. Without access to their system descriptions at the time of writing, we identify three areas where our pipeline could be improved: (1) larger models or parameter-efficient fine-tuning (LoRA) on bigger NLLB variants; (2) higher-quality augmentation validated by LSM experts or via back-translation; (3) constrained decoding or post-processing to enforce gloss conventions in Subtask B.

6.5. Error Analysis: The Gloss Generation Problem

The main failure mode in Subtask B is the model generating fluent Spanish instead of glosses. This happens because both directions use `spa_Latn` as the language token, and with only 968 fine-tuning examples, the pretrained prior toward grammatical Spanish is not always overridden. Three mitigations are worth testing: using an unused NLLB language code for the gloss side, applying constrained decoding to force upper-case output, and increasing augmented data volume.

7. Conclusions and Future Work

We built a pipeline for LSM–Spanish bidirectional translation, testing model selection and data augmentation decisions empirically before the final submission. The system placed 9th on both subtasks (Task Score 0.3724 for MSLG→SPA, 0.1003 for SPA→MSLG).

Three findings emerged from the pipeline:

1. **Architecture choice dominates:** NLLB-200-600M outperformed mT5-small by $4\times$ on BLEU, confirming that translation-specific pretraining matters more than model generality in this setting.
2. **LLM augmentation is practical but limited:** Doubling the dataset to 968 pairs helped, but synthetic pairs do not fully capture LSM gloss conventions (positional classifiers, dual pronouns).
3. **Gloss generation is the harder direction:** The model occasionally outputs Spanish instead of glosses in Subtask B, because the `spa_Latn` language token does not distinguish the two output formats.

Future work should explore larger NLLB variants with LoRA, constrained decoding for Subtask B, domain-specific tokenizers for gloss conventions, and validation of augmented data with LSM experts.

Acknowledgments

We thank the MSLG-SPA 2026 organizers for creating this shared task. Computational resources were provided by Google Colab. Data augmentation used the Claude API (Anthropic).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: data augmentation (synthetic gloss–Spanish pair generation) and grammar review. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[2] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural Sign Language Translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7784–7793.

[3] A. Y. Rodríguez-González, M. Á. Fajardo-Delgado, Daniel and Álvarez-Carmona, M. G. Sánchez-Cervantes, Á. Díaz-Pacheco, R. Aranda, S. Fajardo-Flores, C. H. Martínez-Rodríguez, Overview of mslg-spa at iberlef 2026: Bidirectional translation between mexican sign language glosses and spanish, volume 77, 2026.

[4] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Baez, L. Bali, N. Barrier, S. Bber, I. Birber, M. Brusber, et al., No Language Left Behind: Scaling Human-Centered Machine Translation, arXiv preprint arXiv:2207.04672 (2022).

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021) 483–498.

[6] Anthropic, Claude: A Family of Large Language Models, https://www.anthropic.com/claude, 2025. Accessed: 2026.

[7] K. Yin, J. Read, Better Sign Language Translation with STMC-Transformer, in: Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020, pp. 5975–5989.

[8] M. De Coster, K. D'Oosterlinck, M. Pizurica, P. Rabaey, S. Verlinden, M. Van Herreweghe, J. Dambre, Frozen Pretrained Transformers for Neural Sign Language Translation, in: Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), 2021, pp. 88–97.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[10] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016, pp. 86–96.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2002, pp. 311–318.

[12] S. Banerjee, A. Lavie, 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, 2005, pp. 65–72.

[13] M. Popović, chrF: Character n-gram F-score for Automatic MT Evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2015, pp. 392–395.

[14] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A Neural Framework for MT Evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 2685–2702.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45.

[17] N. Shazeer, M. Stern, Adafactor: Adaptive Learning Rates with Sublinear Memory Cost, in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 4596–4604.