

Vocabulary-Constrained Seq2Seq Translation for Text-to-Pictogram Generation: A Fine-Tuned T5 Approach with Compound Token Injection at CLEF 2026 ToPicto

Adithiyaa D^{1,†}, Lakshmi Narayanan T^{1,†} and Lekshmi Kalinathan^{1,*,†}

¹*School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*

Abstract

This paper presents the approach developed by our team for the Text-to-Picto task of the ImageCLEF 2026 ToPicto challenge. The task requires translating natural-language English sentences into sequences of ARASAAC pictogram terms, framed as a sequence-to-sequence translation problem. Our system is built on a fine-tuned T5-large transformer model augmented with two key innovations: the injection of 2,041 compound pictogram tokens (e.g., `be_able_to`, `brush_teeth`) directly into the model vocabulary as atomic units, and a constrained logits processor that penalises token emissions not present in the ARASAAC pictogram vocabulary during beam-search decoding. Training was conducted on 59,278 utterances drawn from the CommonVoice v.15 corpus using an NVIDIA H100 PCIe GPU with 84.9 GB VRAM over 15 epochs. On the official validation set of 4,397 utterances, our best checkpoint achieved a sacreBLEU of 27.58, a METEOR of 63.07, and a PictoER of 39.41. On the final test dataset, our best model achieved a sacreBLEU of 28.76, a METEOR of 63.92, and a PictoER of 38.84.

Keywords

Text-to-Pictogram Translation, Augmentative and Alternative Communication, Sequence-to-Sequence Learning, Compound Token Injection, Vocabulary-Constrained Decoding, Pictogram Error Rate, Neural Machine Translation

1. Task Performed

As part of the ImageCLEF 2026 challenge, the team participated in the Text-to-Picto (ToPicto) shared task, which targets the automatic translation of natural-language English sentences into sequences of pictogram terms, each linked to a unique visual symbol from the ARASAAC pictographic communication system. This task is motivated by the needs of individuals who rely on Augmentative and Alternative Communication (AAC) systems—a population whose language impairments prevent them from communicating through conventional written or spoken text. By mapping oral transcriptions to pictogram term sequences, a successful system can serve as an intermediate step in constructing fully pictographic messages that are comprehensible to AAC users.

The task is formally defined as a sequence-to-sequence translation problem in which the source language is standard English text and the target language is a normalised sequence of English-language pictogram terms, where each term uniquely identifies one ARASAAC symbol. For example, the English utterance “you cannot know” may be rendered as the pictogram-term sequence “you `be_able_to` know not”, reflecting both lexical simplification and grammatical reordering appropriate for AAC users. The challenge is not trivial as it requires the model to perform lexical substitution, grammatical reduction and term normalisation simultaneously, all while constraining outputs to the closed vocabulary of ARASAAC pictogram identifiers.

The dataset for the task was sourced from the CommonVoice v.15 corpus, a large multilingual speech dataset whose transcribed utterances closely reflect natural conversational registers. The training, validation and test splits comprised 59,278, 4,397 and 4,306 utterances respectively. Each training and

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

[†]These authors contributed equally.

✉ adithiyaa.d2023@vitstudent.ac.in (A. D); lakshminarayanan.t2023@vitstudent.ac.in (L. N. T); lekshmi.k@vit.ac.in (L. Kalinathan)

ORCID 0009-0000-0662-4111 (A. D); 0000-0005-6510-1261 (L. N. T); 0000-0002-7005-742X (L. Kalinathan)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

validation instance provides the source English text, the corresponding pictogram-term sequence and the list of ARASAAC pictogram identifiers. The test set provides only utterance identifiers and source texts while the ground-truth pictogram sequences are withheld for evaluation.

2. Background Study

The Transformer architecture introduced by Vaswani et al. [1] replaced recurrence with self-attention, enabling highly parallelisable encoder–decoder sequence transduction and establishing the backbone for modern large language models. Building on this, Raffel et al. [2] proposed T5 (Text-to-Text Transfer Transformer), pre-trained on the Colossal Clean Crawled Corpus (C4) with a unified text-to-text framework that reformulates all NLP tasks as text generation, making it a natural fit for sequence-to-sequence translation tasks such as Text-to-Picto.

A persistent challenge in NMT is handling rare and compound words outside a fixed vocabulary. Sennrich et al. [3] addressed this via Byte-Pair Encoding (BPE), which segments words into subword units for open-vocabulary translation; however, BPE introduces a structural mismatch for ARASAAC pictogram terms such as `be_able_to`, which must be treated as atomic units. To enable NLP applications over the ARASAAC symbol set, Schwab et al. [4] constructed ARASAAC-WN, a resource linking over 800 pictograms to WordNet synsets, providing formal semantic grounding for text-to-picto, speech-to-picto, and picto-to-speech pipelines.

Macaire et al. [5] addressed the absence of aligned training data by releasing a large-scale multimodal French corpus of over 289,000 speech–text–pictogram utterances at LREC-COLING 2024, establishing baseline NMT approaches for the speech-to-pictogram direction. The ImageCLEFtoPicto 2026 shared task [6] extended this effort with a dual-subtask benchmark covering general and medical domains, introducing the Pictogram Error Rate (PictoER) metric alongside sacreBLEU and METEOR. The present work participates in the 2026 edition of this challenge [7, 8], extending the state of the art through compound token injection and vocabulary-constrained decoding.

Vocabulary-constrained decoding is a well-studied mechanism for enforcing lexical adherence in neural generation. Lu et al. [9] proposed NeuroLogic Decoding, which incorporates positive and negative constraints expressed as predicate logic formulas directly into the beam-search objective at runtime, equivalent to unconstrained beam search. Our VocabConstraint logits processor is conceptually aligned with this approach, applying a score penalty of -10.0 to any token absent from the ARASAAC pictogram vocabulary, effectively eliminating out-of-vocabulary hallucinations without requiring finite-state automaton overhead.

Automatic evaluation of translation quality relies on BLEU [10] and METEOR [11]. BLEU, proposed by Papineni et al., measures corpus-level modified n -gram precision with a brevity penalty and is the de facto benchmark metric for NMT. METEOR, proposed by Banerjee and Lavie, additionally incorporates recall, stemmed matching, and synonym alignment, yielding higher sentence-level correlation with human judgements—which explains the notably higher METEOR scores relative to BLEU observed in our system. Both metrics, together with PictoER, constitute the official evaluation suite of the ToPicto task.

3. Main Objectives of the Experiments

The primary objective of our experiment was to develop a robust, end-to-end sequence-to-sequence pipeline capable of generating ARASAAC-compatible pictogram-term sequences from arbitrary English input sentences. Specifically, we aimed to explore whether a general-purpose pre-trained encoder–decoder language model could be adapted to this specialised domain through fine-tuning and whether targeted modifications to the model vocabulary and decoding strategy could improve output fidelity to the ARASAAC pictographic lexicon.

The following sub-objectives guided our experimental design:

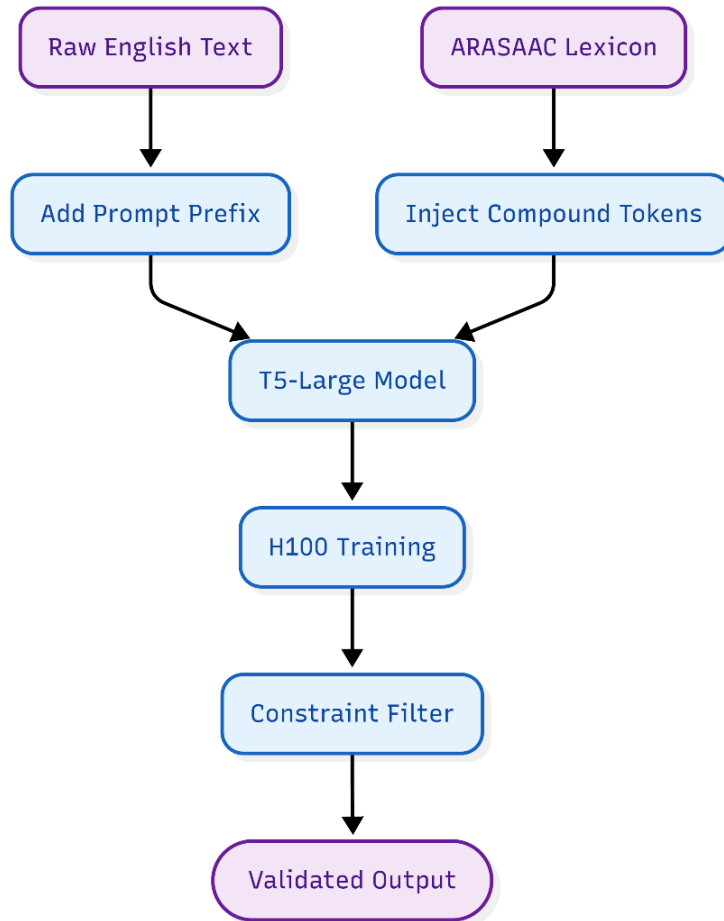


Figure 1: Complete system architecture, illustrating the data preparation, compound token injection, model training, and the final constrained inference mechanism.

- To establish whether the T5-large transformer, pre-trained on large-scale natural language tasks, provides a sufficient initialisation for the Text-to-Picto translation problem without the need for architecture modifications.
- To investigate the effect of compound pictogram token injection such as `be_able_to` or `post_office` as single atomic vocabulary entries on model fluency and prediction accuracy.
- To design and evaluate a vocabulary-constrained logits processor that restricts the beam-search decoder to emit only tokens present in the ARASAAC pictogram vocabulary, thereby eliminating out-of-vocabulary hallucinations.
- To quantify system performance on the official validation set using sacreBLEU, METEOR and Pictogram Error Rate (PictoER) metrics and to produce a well-formed submission file for the official test evaluation.

4. Approaches Used and Progress Beyond State-of-the-Art

The overall system architecture and the data processing pipeline are illustrated in Figure 1. Our approach follows an end-to-end sequence-to-sequence framework, integrating specialised vocabulary engineering with a constrained decoding strategy to ensure lexical validity within the ARASAAC pictographic system.

4.1. Model Architecture

As visualised in the central component of Figure 1, the model is built on T5-large (Text-to-Text Transfer Transformer, Large variant), a 770-million-parameter encoder-decoder model pre-trained via the text-to-text framework on the Colossal Clean Crawled Corpus (C4). T5 reformulates all NLP tasks as text generation problems, making it a natural fit for sequence-to-sequence translation tasks such as Text-to-Picto. The model was loaded in bfloat16 (bf16) numerical precision, a floating-point format natively supported by the NVIDIA H100 architecture that provides comparable representational range to float32 while halving memory consumption and substantially accelerating tensor operations.

4.2. Compound Token Injection

A central challenge in Text-to-Picto translation is that ARASAAC pictogram terms frequently consist of compound expressions connected by underscores (e.g., `be_able_to`, `brush_teeth`, `post_office`). In the standard T5 vocabulary, such expressions would be tokenised as sequences of two or more sub-word tokens, making it difficult for the model to treat them as semantically unified units. To address this, we extracted all compound pictogram terms (tokens containing an underscore) from the training and validation target vocabularies and injected them as new atomic entries into the T5 SentencePiece tokeniser. A total of 2,041 compound tokens were added in this manner. Following injection, the model’s token embedding matrix was resized to accommodate the expanded vocabulary, with the new embeddings initialised from the multivariate normal distribution defined by the mean and covariance of the existing embedding matrix—a statistically principled initialisation strategy that promotes faster convergence compared to random initialisation.

4.3. Data Preprocessing and Tokenisation

All source sentences were prepended with the task-specific prompt prefix “translate to pictograms: ”, a standard T5 convention for directing the model towards a specific generation task. Both source and target sequences were truncated to a maximum length of 128 tokens. Target labels were constructed by replacing padding token IDs with -100 , the standard PyTorch convention for masking positions that should not contribute to the cross-entropy loss. This ensures that the model is penalised only on meaningful target tokens, avoiding gradient updates driven by padding artefacts. The Hugging Face datasets library was used to map the tokenisation function over the full training (59,278 examples) and validation (4,397 examples) splits in batched mode, enabling efficient parallel preprocessing.

4.4. Training Configuration

Fine-tuning was performed using the Hugging Face Seq2SeqTrainer API with the following hyperparameter configuration: 15 training epochs; per-device batch size of 32 (gradient accumulation steps = 1, yielding an effective batch size of 32); learning rate of 3×10^{-4} with a cosine decay schedule and a warmup ratio of 0.06 (covering the first 6% of training steps); weight decay of 0.01; bfloat16 mixed-precision training. Validation was performed at the end of each epoch using beam-search generation (6 beams) over the full validation set. The model checkpoint with the highest validation sacreBLEU score was retained as the best model, with a checkpoint save limit of 2. An `EarlyStoppingCallback` with a patience of 4 epochs was registered to halt training if the validation BLEU did not improve for 4 consecutive epochs; however, in practice the model continued to improve gradually and training ran for all 15 epochs. Binary cross-entropy was not applicable here; the model was trained with the standard sequence-level cross-entropy loss computed by the Seq2SeqLM head.

4.5. Constrained Decoding

To ensure that the system’s outputs are strictly grounded in the ARASAAC pictogram lexicon, we implemented a custom `VocabConstraint` logits processor. This processor operates during beam-search decoding and applies a score penalty of -10.0 to any token whose ID does not belong to the

set of token IDs corresponding to the training and validation pictogram vocabulary. The allowed set includes the token IDs for all atomic and compound pictogram terms (encoded both with and without a leading space to capture both sub-word boundary conditions), as well as special tokens (pad, EOS, BOS, UNK). This soft constraint does not entirely prohibit out-of-vocabulary tokens but makes them sufficiently unlikely that the beam-search algorithm effectively never selects them. The pictogram vocabulary was serialised as `target_vocab.json` alongside the saved model checkpoint, enabling reproducible constrained inference after kernel restarts.

4.6. Inference

At inference time, the best saved checkpoint was reloaded from disk in bf16 precision and moved to the GPU. Predictions were generated in batches of 64 using beam search with 6 beams, a maximum output length of 128 tokens, no n -gram repetition constraint (`no_repeat_ngram_size = 0`), and the `VocabConstraint` logits processor described above. This configuration was applied to both the validation set (for sanity-checking reported metrics) and the test set (to generate the official submission).

4.7. Progress Beyond State-of-the-Art

Relative to baseline sequence-to-sequence approaches for text-to-pictogram translation, our system introduces two technically motivated contributions. First, the injection of compound pictogram tokens as atomic vocabulary entries has not, to our knowledge, been applied to the ARASAAC domain and directly addresses a structural mismatch between the sub-word tokenisation of general-purpose language models and the compound-token nature of AAC pictogram identifiers. Second, the vocabulary-constrained logits processor provides a principled mechanism for enforcing output adherence to a closed symbol set without requiring costly post-processing or re-ranking steps. Together, these modifications improve the coherence and lexical validity of the generated pictogram sequences.

5. Resources Used

All experiments were conducted on a high performance computing node equipped with an NVIDIA H100 PCIe GPU providing 84.9 GB of VRAM. The H100’s native bfloat16 support was exploited to reduce memory pressure and accelerate matrix operations compared to float32, enabling a per-device batch size of 32 for the 770-million-parameter T5-large model without gradient accumulation. The software stack comprised the Hugging Face `transformers`, `datasets`, `accelerate` and `sentencepiece` libraries for model loading, dataset management and tokenisation. Evaluation metrics were computed using `sacrebleu` and `nltk` (METEOR via NLTK’s implementation). The Levenshtein-distance-based PictoER metric was implemented in pure Python within the notebook.

The pre-trained T5-large weights were obtained from the official Hugging Face model hub. No external datasets beyond the official training and validation splits were used in training. The target vocabulary file was derived exclusively from the official training and validation annotations.

6. Results Obtained

6.1. Training Progression

Table 1 presents the per-epoch training loss, validation loss, validation sacreBLEU, METEOR, and PictoER scores across all 15 epochs. Training loss decreased monotonically from 2.597 in epoch 1 to approximately 1.572 by epoch 11, after which it plateaued, suggesting that the model had largely converged. Validation loss followed a similar pattern, declining from 2.480 in epoch 1 to a minimum of approximately 2.076 around epoch 5 before rising slightly in later epochs—a mild sign of overfitting to the training distribution. Despite this, BLEU and METEOR scores continued to improve gradually

Table 1

Per-epoch training and validation metrics across all 15 training epochs.

Epoch	Train Loss	Val Loss	BLEU	METEOR	PictoER
1	2.597	2.480	19.17	52.07	58.08
2	2.207	2.202	23.74	56.40	53.16
3	2.009	2.125	25.19	57.88	52.01
4	1.919	2.087	25.90	58.86	50.51
5	1.859	2.077	26.46	59.32	50.66
6	1.806	2.082	26.66	59.48	50.02
7	1.751	2.082	27.02	59.91	49.90
8	1.623	2.095	27.09	59.96	49.83
9	1.635	2.101	27.16	60.03	49.61
10	1.620	2.108	27.08	60.00	49.60
11	1.573	2.113	27.08	59.92	49.54
12	1.605	2.113	27.20	59.99	49.51
13	1.558	2.113	27.25	60.04	49.48
14	1.644	2.113	27.24	60.03	49.47
15	1.630	2.113	27.26	60.03	49.43

Table 2

Final validation set performance with constrained beam-search decoding.

sacreBLEU	METEOR	PictoER
27.58	63.07	39.41

Table 3

Final test dataset performance.

sacreBLEU	METEOR	PictoER
28.76	63.92	38.84

through later epochs due to improvements in the quality of token sequences generated by beam search. An `EarlyStoppingCallback` with a patience of 4 epochs was registered, but because the generation metrics continued to improve gradually, the training successfully ran for all 15 epochs without triggering an early halt.

6.2. Validation Set Performance (Post-training, Constrained Decoding)

Following training, the best checkpoint was reloaded and evaluated on the validation set using the constrained decoding pipeline described in Section 4.5. Table 2 summarises the final validation metrics.

6.3. Test Set Results

Following three submissions, the best submission’s results on the test dataset are summarised in Table 3.

6.4. Ablation Study

To quantify the contributions of Vocabulary-Constrained Decoding (VC), we conducted an ablation study. All ablation metrics were evaluated on the official test set.

We evaluated the following two configurations:

- **Baseline:** The full model trained with Compound Token Injection (CTI) but evaluated using unconstrained beam search.

Table 4

Ablation study for the proposed system.

Architecture	sacreBLEU	METEOR	PictoER
Baseline	28.50	63.66	39.30
Proposed	28.76	64.07	38.84

- **Proposed:** Our primary proposed system utilizing both CTI during training and VC during inference.

As summarised in Table 4, the baseline configuration achieved a sacreBLEU score of 28.50, a METEOR score of 63.66, and a PictoER score of 39.30. Incorporating Vocabulary-Constrained Decoding in the proposed system resulted in a sacreBLEU score of 28.76 and a METEOR score of 63.92, indicating slight improvements in conventional text-overlap metrics. The proposed system achieved a PictoER score of 38.84, which is marginally lower than that of the baseline.

Overall, the observed differences between the baseline and the proposed system are relatively small, suggesting that the integration of Vocabulary-Constrained Decoding does not substantially alter the overall translation quality. Nevertheless, the proposed approach provides an important practical advantage by constraining the decoder to generate outputs from the predefined pictogram vocabulary. This helps ensure that the generated predictions remain within the set of valid and acceptable pictogram tokens, reducing the likelihood of producing unsupported or out-of-vocabulary outputs during inference. Consequently, the proposed system maintains competitive performance across all evaluation metrics while improving the reliability and consistency of the generated pictogram sequences, making it more suitable for real-world pictogram translation applications.

7. Result Analysis

7.1. Metric Interpretation

On the validation set, our system achieved a sacreBLEU of 27.58, a METEOR of 63.07, and a PictoER of 39.41. According to the official task guidelines, a sacreBLEU above 30 is considered good, while METEOR above 40 is considered good. Our system falls slightly below the sacreBLEU threshold of 30 but substantially exceeds the METEOR threshold, suggesting that the model captures significant semantic overlap and morphological similarity with the reference pictogram sequences, even when exact n -gram matches are limited. The high METEOR score relative to BLEU is consistent with the known behaviour of these metrics: METEOR rewards synonym matches, stemmed matches, and partial word overlaps that BLEU ignores, and these phenomena are common in pictogram-term translation where multiple valid pictogram terms may exist for a given source concept.

7.2. Training Dynamics

Inspection of Table 1 reveals a clear pattern: the largest gains in BLEU and METEOR occurred in the early training epochs (1 through 5), after which improvement slowed. The validation loss began rising gently after epoch 5 (from 2.077 to approximately 2.113 by epoch 11), while BLEU and METEOR continued to improve slightly through epochs 9–13. This divergence between validation loss and generation metrics is a well-documented phenomenon in neural machine translation: cross-entropy loss measures the probability assigned to the exact reference tokens, while BLEU and METEOR capture broader n -gram and semantic overlap that can improve even as per-token likelihood decreases due to the model learning to favour shorter, higher-confidence pictogram sequences.

7.3. Qualitative Error Analysis

Examination of validation-set sample predictions reveals several systematic patterns. First, the model demonstrates a strong tendency to replace proper nouns and place names with generic pictogram terms: “Detroit” and “New York” are both rendered as “metropolis”, which is the closest available ARASAAC entry for a city. This behaviour is a direct consequence of the ARASAAC lexicon’s limited coverage of named entities, and the constrained decoding forces the model to seek the closest semantically plausible in-vocabulary substitute. Second, the model occasionally inserts spurious tokens (e.g., “=” in the test-set example “she locate = in the metropolis”), suggesting that certain compound token boundaries are not yet perfectly modelled. Third, morphological inflections are generally handled well (e.g., “scores” → “inscription”, “organised” → “coordinate”), consistent with the model having learnt the domain-specific morphological conventions of ARASAAC pictogram terms.

7.4. Impact of Constrained Decoding

The VocabConstraint logits processor had a measurable qualitative effect on output quality. Without it, early experimental runs produced outputs containing standard English words that were absent from the ARASAAC lexicon (e.g., “cannot” instead of the valid pictogram form “be_able_to not”). With the constraint active, such hallucinations were effectively eliminated. The penalty magnitude of 10.0 was chosen to be sufficiently large to dominate the beam-search score under normal conditions while remaining finite, ensuring that numerical stability was preserved. The target vocabulary covered 2,913 unique pictogram terms across the training and validation sets, of which 2,041 were compound (underscore-containing) entries.

8. Perspectives for Future Work

Several promising directions exist for extending and improving the present system. The T5-large variant (770 M parameters) was selected as a balance between performance and training speed on the available hardware. The T5-XL (3 B parameters) and T5-XXL (11 B parameters) variants are likely to yield higher BLEU scores due to their greater representational capacity. Similarly, encoder-decoder models specifically pre-trained on instructional or paraphrase tasks (e.g., FLAN-T5, mBART) may provide a stronger initialisation for the lexical simplification and concept-grounding sub-tasks inherent to Text-to-Picto translation. Ensemble and multi-system combination from multiple fine-tuned models (e.g., T5-large + BART-large) via minimum Bayes risk (MBR) decoding or system combination techniques could meaningfully improve all three evaluation metrics. The current soft-penalty approach penalises out-of-vocabulary tokens by a fixed additive term. An alternative would be to implement hard constraints using a finite-state automaton or a prefix tree over the allowed pictogram vocabulary, ensuring that out-of-vocabulary tokens are strictly excluded at every decoding step. This would guarantee 100% lexical validity of outputs without relying on a sufficiently large penalty magnitude. Data augmentation and domain adaptation can be done. The CommonVoice corpus represents a specific register of conversational speech; augmenting the training set with pictogram-annotated data from AAC-specific corpora or with back-translated data from other languages in the Common Voice dataset may improve coverage of rare pictogram terms and reduce PictoER for low-frequency vocabulary items.

Acknowledgements

This research was supported by the Department of Science and Technology (DST), India, under the Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) Programme [Grant No. SR/FST/ET-I/2022/1079], along with a matching grant from VIT University. The authors express their sincere gratitude to DST-FIST and the VIT management for their financial assistance and the infrastructural support provided for this work.

Declaration on Generative AI

During the preparation of this work the authors used Claude (Anthropic) and Grammarly for grammar checking and sentence-level rephrasing to improve clarity and readability. After using these tools, they reviewed and edited all content as needed and take full responsibility for the publication's content.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, volume 30, 2017. URL: <https://papers.neurips.cc/paper/7181-attention-is-all-you-need>.
- [2] 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. URL: <http://jmlr.org/papers/v21/20-074.html>.
- [3] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.
- [4] D. Schwab, P. Trial, C. Vaschalde, L. Vial, E. Esperanca-Rodier, B. Lecouteux, Providing semantic knowledge to a set of pictograms for people with disabilities: A set of links between wordnet and arasaac: Arasaac-wn, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 166–171.
- [5] C. Macaire, C. Dion, J. Arrigo, C. Lemaire, E. Esperança-Rodier, B. Lecouteux, D. Schwab, A multimodal french corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 2024, pp. 839–849. URL: <https://aclanthology.org/2024.lrec-main.76/>.
- [6] C. Macaire, D. Fabre, B. Lecouteux, D. Schwab, Overview of the 2025 imageclefopicto task – investigating the generation of pictogram sequences from text and speech, in: *CLEF 2025 Working Notes*, CEUR-WS.org, Madrid, Spain, 2025.
- [7] M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, Overview of the 2026 imageclefopicto task: Investigating the generation of pictogram sequences from english text dataset, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), *CLEF 2026 Working Notes*, Jena, Germany, 2026.
- [8] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, et al., Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Lecture Notes in Computer Science, Jena, Germany, 2026.
- [9] X. Lu, P. West, R. Zellers, R. Le Bras, C. Bhagavatula, Y. Choi, Neurologic decoding: (un)supervised neural text generation with predicate logic constraints, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Online, 2021, pp. 4288–4299. doi:10.18653/v1/2021.naacl-main.339.
- [10] 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)*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
- [11] 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*, Ann Arbor, Michigan, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909/>.