

Towards Multimodal Text-to-Pictogram Generation and Evaluation*

Notebook for the ImageCLEF Lab at CLEF 2026

Sami Ul Haq*, Sheila Castilho

School of Applied Languages and Intercultural Studies (SALIS), ADAPT Centre, Dublin City University (DCU), Dublin, Ireland

Abstract

Augmentative and Alternative Communication (AAC) systems, such as those based on pictograms, aim to facilitate communication for individuals who experience difficulties with natural language. The text-to-pictogram generation task involves translating natural language text into sequences of pictograms, with the goal of reducing the communication gap between AAC users and broader society. In this paper, we investigate whether directly incorporating pictogram images during model training improves overall performance. We evaluate our systems using multiple automatic metrics, including n-gram-based, learned, and LLM-based measures. The results show that systems trained on multimodal inputs (text and pictograms) consistently achieve the highest scores across most metrics.

Keywords

Text-to-Pictogram, Multimodal Generation, Augmentative and Alternative Communication, Automatic Evaluation

1. Introduction

Augmentative and Alternative Communication (AAC) encompasses methods designed to facilitate communication for individuals whose conventional language abilities are impaired or absent [1, 2]. These methods provide a critical lifeline, helping individuals effectively convey their needs, express thoughts, and participate more independently in daily life. AAC also has potential applications in establishing communication in disaster response and relief missions [3], as well as in enhancing patient understanding of complex medical instructions [4].

Visual communication is central to AAC. Common examples include sign language for deaf individuals and the use of pictograms, which serve as simplified, iconic representations of concepts or ideas. This approach has been shown to significantly enhance language accessibility and reduce the cognitive load associated with linguistic processing [5]. Several pictogram sets exist, such as PCS, Widgit, and Pictogram. The most widely used is ARASAAC¹, an online resource of over 25,000 pictograms available under a Creative Commons license (BY-NC-SA), funded by the Department of Culture, Sports, and Education of the Government of Aragon, Spain.

Text-to-pictogram translation is a natural language processing (NLP) task that involves automatically converting natural language text into visual sequences, thereby bridging the communication gap between AAC users and those unfamiliar with pictogram-based systems. In this context, the ToPicto task of ImageCLEF 2026 [6, 7] represents a prominent shared task that fosters research in assistive communication technologies through a series of challenges and evaluation labs.

In this paper, we present our participation in the ImageCLEF ToPicto task. Our main focus is to investigate whether incorporating visual pictogram images alongside pictogram labels improves the performance of text-to-pictogram translation. Specifically, we leveraged a visual language model (VLM)

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

*Corresponding author.

✉ sami.haq2@mail.dcu.ie (S. U. Haq); sheila.castilho@dcu.ie (S. Castilho)

🌐 <https://sami-haq99.github.io//> (S. U. Haq)

🆔 0009-0009-9398-0151 (S. U. Haq); 0000-0002-8416-6555 (S. Castilho)



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

¹<https://arasaac.org>

to encode pictogram images and augmented a text-only baseline with visual signals via a linear projection head. We then compared the performance of different systems using commonly used automatic metrics.

We also expand on the challenge of automatic evaluation for pictogram translation, as the metrics are adapted from the machine translation (MT) domain, yet pictogram sequences differ considerably from natural language text. To address this, we employ a multi-metric evaluation setup and conduct a qualitative analysis to compare the performance of these metrics.

2. Background

Natural language to pictogram translation encompasses both text-to-pictogram and speech-to-pictogram systems. The former typically involves training or fine-tuning models to generate sequences of pictogram terms, which are then mapped to pictogram images such as those available on ARASAAC. Speech-to-pictogram systems adopt two approaches: end-to-end, which directly translates speech into pictogram terms, and cascade systems, which first apply speech recognition and then translate the resulting text into pictograms [8].

Earlier work on pictogram-supported communication focused on constructing text-to-pictogram resources by aligning lexical and semantic databases, and on building translation systems that mapped individual tokens to pictograms [9, 10]. More recently, text-to-pictogram systems have largely adopted Transformer architectures [11, 12] and the fine-tuning of pre-trained language models [13, 14]. Elliah et al. [13] fine-tune a GPT-2 model and a Helsinki-NLP MT model (trained on Romance languages) for the Text-to-Picto task, finding that the Helsinki model achieves better performance as measured by PictoER, BLEU, and METEOR scores. Anand et al. [15] follow a similar approach and fine-tune a BERT-based model, reporting results using automatic metrics but without comparison to any baselines. Hjuler and Fabre [12] fine-tune several variants of Google T5 to generate pictogram sequences, reporting that models with a higher number of parameters consistently achieve better performance.

For speech-to-pictogram translation, Vaschalde et al. [16] proposed a system for converting French speech into ARASAAC pictograms by integrating speech recognition with rule-based simplification and disambiguation. More recently, Macaire et al. [17] introduced the first large-scale aligned multimodal corpus (Propicto) and established benchmarks for speech-to-pictogram translation using both cascade and end-to-end architectures. Hjuler and Fabre [12] also fine-tuned Whisper models for speech-to-pictogram translation, reporting that models with a greater number of parameters achieved superior performance.

The evaluation of pictogram translation quality typically combines automatic MT metrics with expert qualitative review. Standard automatic metrics include BLEU [18] for n-gram overlap, METEOR [19] for semantic alignment, and the Pictogram Error Rate (PictoER), which adapts the word error rate specifically for tokenised pictogram sequences [20].

Overall, existing research has primarily focused on the text and speech modalities, with the visual pictogram modality remaining under-explored in the development and evaluation of these systems. Furthermore, automatic evaluation relies heavily on string-based metrics, which, while well-suited to this task given their language independence, low resource requirements, and exact matching criteria, do not account for semantically acceptable variations in pictogram sequences.

3. Data

The dataset for this task is sourced from the CommonVoice v.15 corpus², a collection of speech data recorded by volunteers on the Common Voice platform. The text is drawn from a variety of public-domain sources, including blog posts, books, films, and other speech corpora, and is considered representative of naturalistic interactions involving pictogram-reliant communication. The data is provided in JSON format across three splits: training (59,278 instances), validation (4,397 instances), and test

²<https://commonvoice.mozilla.org/>

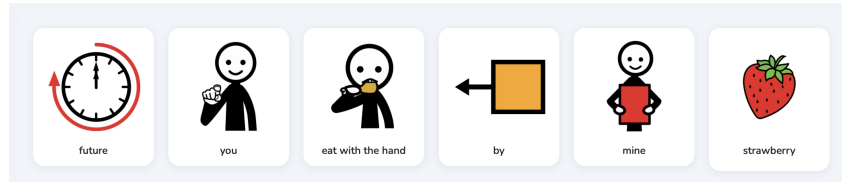


Figure 1: Example of pictogram sequence generated for the source utterance “you will eat some of my strawberries”, showing the corresponding ARASAAC pictograms for each predicted label.

(4,306 instances). Since the official test set was not available during the development phase, we divided the validation set equally to create a self-test set for system development and a held-out validation set for model training.

Table 1 presents an example record from the dataset. The *src* field contains a natural language utterance, while *tgt* and *hyp* are sequences of pictogram labels. All three fields are in English. To visualise the predicted pictogram labels, we developed an online platform³ using the ARASAAC Developer API⁴, which retrieves and displays the corresponding pictogram images, as illustrated in Figure 1.

Table 1
Example record from train split of dataset.

Tag	Definition	Example
id	Unique identifier of each utterance	common_voice_fr_20414147.mp3
src	text from oral transcription	you will eat some of my strawberries
tgt	sequence of pictogram terms	future you eat_with_the_hand by mine strawberry
pictos	List of pictogram identifiers linked to each pictogram term	[9829, 7304, 38413, 7074, 12266, 2400]

4. Text-to-Pictogram Generation

The task of text-to-pictogram translation aims to convert a natural language English sentence into a sequence of pictogram labels that convey the same semantic content in a form compatible with AAC systems. Each label should correspond directly to a symbol in a standardised pictogram library.

Formally, given a source sentence $X = (x_1, x_2, \dots, x_n)$, the model generates a target sequence $Y = (y_1, y_2, \dots, y_m)$ of pictogram labels. The output sequence is typically shorter than the input sentence, as pictogram communication operates at a higher level of abstraction.

We develop and compare two systems: (1) a **text-only** baseline, and (2) a **multimodal** system that incorporates pictogram images alongside textual representations during training.

4.1. Text-Only Baseline and Backbone Model

The text-only system uses T5-small as the sequence-to-sequence backbone model. T5 is trained on a large-scale text corpus, enabling its encoder to learn strong sentence-level semantic representations that reduce the need for extensive labelled data during downstream fine-tuning. The T5-small variant (60M parameters) was selected over larger alternatives to balance representational capacity with the modest size of the available AAC parallel corpus.

³<https://huggingface.co/spaces/samiulhaq/ARASAAC-Pictogram-Viewer>

⁴<https://arasaac.org/developers/api>

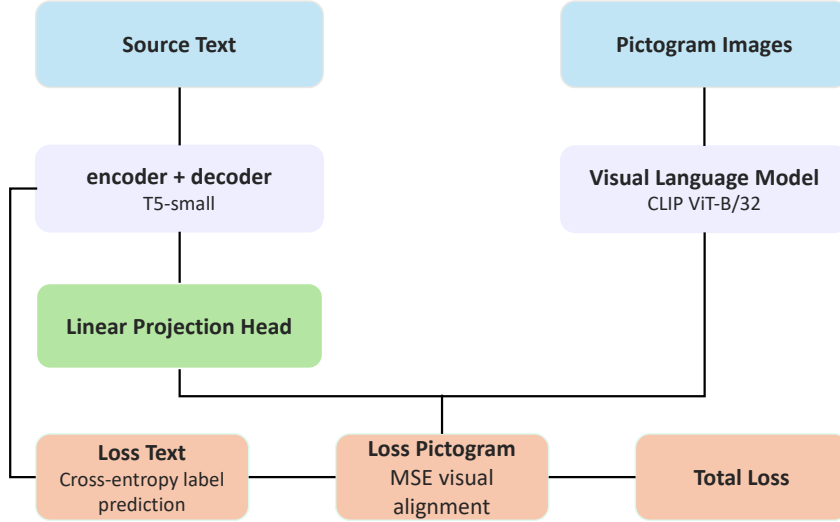


Figure 2: Overview of the text-to-pictogram generation multimodal pipeline.

Source sentences and target pictogram label sequences are tokenised using the T5 tokenizer, with a maximum sequence length of 32 tokens. Model training is performed using a token-level cross-entropy objective over the target label sequence under teacher forcing:

$$\mathcal{L}_{\text{text}} = - \sum_t \log P(y_t \mid y_{<t}, X)$$

The text-only baseline fine-tunes T5-small solely on the parallel corpus of (English sentence, pictogram label sequence) pairs, with no image data at any stage. It is trained end-to-end with $\mathcal{L}_{\text{text}}$ using the AdamW optimiser at a learning rate of 3×10^{-5} and a batch size of 8.

4.2. Multimodal System

The multimodal system augments the text-only baseline with a visual grounding objective that encourages the decoder’s internal representations to align with the visual representations of related pictograms. Unlike arbitrary identifiers, pictogram labels represent concrete visual symbols with a consistent and semantically structured appearance. Grounding the decoder’s representations in the visual space may therefore support better generalisation, particularly for labels whose surface forms are similar but whose meanings differ. Figure 2 presents an overview of the multimodal pipeline for text-to-pictogram translation.

4.2.1. Visual Feature Extraction

Visual features are extracted using CLIP ViT-B/32 [21], a vision-language model pre-trained on 400 million image-caption pairs via a contrastive objective. CLIP serves purely as a fixed feature extractor. During training, the CLIP image encoder is kept frozen to prevent degradation of the visual encoder under the limited number of available pictogram images, and to ensure that the embedding space retains its pre-trained semantic structure.

Since CLIP is frozen, all pictogram embeddings can be pre-computed once before training and cached. Furthermore, since a natural language sentence may correspond to a set of pictograms ($\mathcal{P}$), the embeddings are averaged across all pictograms for each target sequence:

$$\mathbf{v} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{CLIP}_{\text{image}}(\text{img}_p)$$

The text-only system is fine-tuned from T5-small, while the multimodal system incorporates visual features using two variants of CLIP: openai/clip-vit-base-patch32 and openai/clip-vit-large-patch14.

4.2.2. Projection Head and Visual Grounding Loss

A learned linear projection layer (see Figure 2) maps the decoder’s averaged hidden states to CLIP’s embedding space, producing a sequence-level visual representation. Mean pooling over the full decoder output is used rather than token-level alignment, as the correspondence between output tokens and pictograms is variable and non-monotone. The projection head is trained jointly with T5.

The visual grounding loss is the Mean Squared Error (MSE) between the projected decoder representation and the pre-computed CLIP pictogram embedding, computed only over examples with valid visual targets. The total training objective combines both losses, with a weighting coefficient $\lambda = 0.3$ that positions the visual grounding term as a soft regulariser:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \lambda \cdot \mathcal{L}_{\text{picto}}$$

4.3. Training

Both systems are trained using the AdamW optimiser with a learning rate of 3×10^{-5} and a batch size of 8. Early stopping is applied by monitoring the validation loss at the end of each epoch, terminating training if no improvement is observed for three consecutive epochs, at which point the best-performing checkpoint is restored.

Model performance is reported using corpus-level SacreBLEU on generated label sequences under greedy decoding. This metric and evaluation procedure are identical across both systems, ensuring a direct and controlled comparison.

4.4. Evaluation

We employ a multi-metric evaluation setup combining lexical, character-level, and semantic similarity metrics. Pictogram translations typically produce short, semi-structured token sequences with a limited vocabulary and frequent symbolic variations, making reliance on a single metric insufficient. We report the performance of text-to-pictogram translation systems in Section 4.5, using following metrics:

BLEU: evaluates n-gram precision between candidate and reference sequences, and has been widely used in MT and sequence generation tasks. We include BLEU primarily to enable comparison with prior work, rather than as a principal indicator of translation accuracy.

chrF [22]: a character n-gram F-score metric that is more robust than BLEU in morphologically rich or low-resource settings. chrF is particularly suitable for our task, as pictogram label outputs often contain minor token variants such as inflectional differences (e.g., *create* vs. *creates*). Operating at the character level, it provides partial credit for such variations and improves sensitivity to near-matches in short sequences.

TER [23]: Translation Edit Rate measures the minimum number of edit operations — insertions, deletions, substitutions, and shifts — required to transform a system output into its reference. In our setting, TER serves as a complementary measure of structural deviation and post-editing effort.

BERTScore [24]: unlike surface-based metrics, BERTScore computes similarity using contextualised token embeddings from pre-trained transformer models, capturing semantic equivalence even when surface forms differ. This is particularly relevant in our task, where multiple surface forms may correspond to similar pictographic meanings.

Qwen-2.5: Finally, we employ Qwen2.5-72B-Instruct⁵ as an LLM-as-Judge metric. This model is instruction-tuned across over 29 languages and supports contexts of up to 128K tokens. We select it for its broad linguistic coverage, large context window, and open-source availability, which together make it well-suited to assess pictogram translation quality beyond simple string matching. The prompt used for LLM-based evaluation of text-to-pictogram translation is provided in Appendix A.

4.5. Results

Table 2 presents the Type-Token Ratio (TTR) of different system outputs and the reference text from the test set. The reference text achieves the highest TTR, scoring 13 percentage points (2913 unique types) higher than the T5-CLIP_{large} output, which in turn scores 11 percentage points (2603 unique types) higher than the text-only system.

Table 2

Type-Token Ratio (TTR) comparison of system outputs and reference text.

Category	System	Tokens	Types	TTR
Text-only	T5 _{small}	27,126	2,590	0.0955
Multimodal	T5-CLIP _{base}	26,923	2,563	0.0952
Multimodal	T5-CLIP _{large}	26,938	2,603	0.0966
Reference	Official	27,240	2,913	0.1069

4.6. System-level results

Table 3 reports system-level results across both test sets. Overall, the scores are broadly similar across systems, with differences of one to two points in most metrics. T5_{small} and T5-CLIP_{large} are the closest competitors, while T5-CLIP_{base} shows a more noticeable gap across most metrics. The final two columns of Table 3 present the Adequacy and Faithfulness scores assigned by the LLM judge (Qwen2.5) on the official test set. The performance gap remains narrow: T5-CLIP_{large} leads on Adequacy, while T5-CLIP_{base} achieves the highest Faithfulness score, with all differences smaller than one point.

Table 3

Evaluation results on Self and Official test sets. Best scores per column within each test set are highlighted in bold.

Test Set	Model	BLEU ↑	TER ↓	chrF ↑	BERTScore ↑	Qwen _{adeq} ↑	Qwen _{faith} ↑
Self	T5 _{small}	49.44	33.67	64.15	0.8868	-	-
	T5-CLIP _{Base}	47.99	34.65	62.78	0.8835	-	-
	T5-CLIP _{Large}	49.70	33.29	64.26	0.8868	-	-
Official	T5 _{small}	51.12	32.94	65.27	0.8895	3.683	2.978
	T5-CLIP _{Base}	49.47	34.07	63.84	0.8852	3.635	2.987
	T5-CLIP _{Large}	51.33	32.37	65.40	0.8899	3.693	2.982

4.6.1. LLM vs. MT Metrics

In addition, we examine how automatic evaluation metrics (both n-gram- and embedding-based) correlate with LLM-based judgments at the sentence level. Since human evaluation was not available for this dataset, we use LLM judgments as a proxy for semantic assessment, given the strong language understanding and reasoning capabilities of these models. Table 4 presents the Spearman correlation coefficients between automatic metrics and LLM-based assessments. Among the evaluated metrics, BERTScore (using recall) exhibits the strongest positive correlation with LLM judgments, while chrF also demonstrates competitive performance.

⁵<https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

Table 4

Spearman correlations between LLM evaluation scores and MT metrics across systems.

Metric	T5_{small}	T5-CLIP_{Base}	T5-CLIP_{Large}
ChrF4	0.7258	0.7203	0.7349
ChrF4W2	0.7128	0.7033	0.7176
ChrF6	0.7137	0.7074	0.7227
ChrF6W2	0.7076	0.6985	0.7135
BLEU	0.5989	0.5829	0.5985
TER	-0.6624	-0.6492	-0.6685
BERTScore_P	0.7287	0.7199	0.7305
BERTScore_R	0.7562	0.7452	0.7598

4.6.2. Qualitative Analysis

To further investigate and analyse the results, we manually inspected the output of text-to-pictogram models and evaluation metrics and observed that each evaluation approach presents distinct challenges, including the LLM-as-a-judge paradigm (see Appendix B and Table 5). For instance, both BERTScore and LLM-based evaluations tend to assign higher scores to hypotheses that share similar concepts with the reference, occasionally conflating semantic relatedness with overall translation quality. In contrast, n-gram-based metrics such as BLEU and chrF often penalise acceptable lexical or structural variations in the hypothesis, a well-documented limitation of this class of metrics.

Interestingly, because text-to-pictogram sequences are typically short and the target labels are drawn from a predefined pictogram vocabulary, chrF with character 4-grams demonstrated relatively better performance during manual inspection and achieved competitive correlations with LLM-based evaluations⁶. This observation is consistent with previous findings that highlight the robustness of chrF on novel evaluation tasks [25].

Regarding the LLM-based evaluations, it is worth noting that we initially designed the evaluation prompts around two dimensions: adequacy and faithfulness. The source text, reference, and hypothesis were provided within a single prompt. However, we observed that the LLM often produced unreliable evaluations despite access to the reference text. This issue appears to stem from the presence of natural language source text in English alongside text-to-pictogram labels, which do not follow conventional grammatical structures and frequently contain multiple words joined by hyphens (e.g., “at-the”). As a result, the model struggled to consistently follow the evaluation instructions and distinguish between the different inputs.

After several rounds of experimentation, we removed the source text from the prompt and restricted the evaluation to a single dimension, adequacy (see the final prompt in Appendix A). Our findings suggest that simpler prompts focusing on one evaluation dimension at a time, supplemented with a small number of task-specific examples, lead to more reliable and consistent LLM-based evaluations.

5. Conclusion

In this paper, we investigated whether incorporating visual pictogram representations into model training improves text-to-pictogram translation. We trained three systems: a text-only T5_{small} baseline and two multimodal variants (T5-CLIP_{Base} and T5-CLIP_{Large}) that incorporate CLIP visual features. Automatic evaluation showed that the multimodal systems achieved the best scores across most metrics, although the differences were small, suggesting a modest rather than substantial benefit from visual grounding.

Qualitative analysis further highlighted the limitations of automatic metrics. In particular, we found that chrF4 provides more balanced scores and correlates more strongly with LLM-based evaluation at the segment level. A key limitation of this work is that human evaluation was not performed,

⁶Although LLM-based judgments cannot be considered a gold standard, they provide a useful baseline in the absence of human evaluation due to their ability to capture semantic relationships and contextual meaning.

which would have provided more reliable assessments. In future work, we plan to incorporate human evaluation and develop multimodal automatic metrics to further validate the role of visual information in text-to-pictogram generation.

Acknowledgments

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. This research was also supported by Research Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT Centre at Dublin City University. ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology, is funded through the Research Ireland Centres Programme. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Declaration on Generative AI

During the preparation of this work, the author used Overleaf's built-in AI assistant for grammar and spelling checks. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] J. Light, D. McNaughton, J. Caron, New and emerging aac technology supports for children with complex communication needs and their communication partners: State of the science and future research directions, *Augmentative and Alternative Communication* 35 (2019) 26–41.
- [2] M. Ronski, R. A. Sevcik, Augmentative communication and early intervention: Myths and realities, *Infants & Young Children* 18 (2005) 174–185.
- [3] T. H. Clawson, J. Leafman, G. M. Nehrenz Sr, S. Kimmer, Using pictograms for communication, *Military medicine* 177 (2012) 291–295.
- [4] R. Malhotra, M. A. C. Bautista, N. C. Tan, W. E. Tang, S. Tay, A. S. L. Tan, A. Pouliot, S. E. Saffari, C.-L. Chei, R. Vaillancourt, Bilingual text with or without pictograms improves elderly singaporeans' understanding of prescription medication labels, *The Gerontologist* 59 (2019) 378–390.
- [5] R. Hervás, S. Bautista, G. Méndez, P. Galván, P. Gervás, Predictive composition of pictogram messages for users with autism, *Journal of Ambient Intelligence and Humanized Computing* 11 (2020) 5649–5664.
- [6] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [7] M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, Overview of the 2026 imagecleftopicto task: Investigating the generation of pictogram sequences from english text dataset, in: *CLEF 2026 Working Notes, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Jena, Germany, 2026.

- [8] C. Macaire, C. Dion, D. Schwab, B. Lecouteux, E. Esperança-Rodier, Towards speech-to-pictograms translation, in: *Interspeech 2024*, 2024, pp. 857–861.
- [9] R. Mihalcea, C. W. Leong, Toward communicating simple sentences using pictorial representations, *Machine translation* 22 (2008) 153–173.
- [10] V. Vandeghinste, I. Schuurman, L. Sevens, F. Van Eynde, Translating text into pictographs, *Natural Language Engineering* 23 (2017) 217–244.
- [11] A. Koushik, J. Morrison S., P. Mirunalini, J. A. R. K., A transformer based approach for text-to-picto generation, in: *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, 2024.
- [12] M. J. Hjuler, I. Fabre, Parlez-vous picto? a transformer-based approach for text-to-picto and speech-to-picto translation in french, in: *CLEF 2025 Working Notes*, CEUR Workshop Proceedings, 2025.
- [13] A. Elliah, A. Narayanan P., S. Bhuvan, P. Mirunalini, Text-to-picto using lexical simplification, in: *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, 2024.
- [14] C. Macaire, D. Fabre, B. Lecouteux, D. Schwab, Overview of the 2025 imagecleftpicto task—investigating the generation of pictogram sequences from text and speech, in: *CLEF2025 Working Notes*. CEUR Workshop Proceedings, CEUR-WS. org, Madrid, Spain, 2025.
- [15] B. Anand, J. Themozhi, S. Sai R., P. Charumathi, T. T. Mirnalinee, Ssn-mlrg at text to picto 2024: A bert-based approach for mapping french sentences to pictogram terms, in: *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, 2024.
- [16] C. Vaschalde, P. Trial, E. Esperança-Rodier, D. Schwab, B. Lecouteux, Automatic pictogram generation from speech to help the implementation of a mediated communication, in: *Conference on Barrier-free Communication*, 2018.
- [17] 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)*, ELRA and ICCL, 2024, pp. 839–849.
- [18] 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*, 2002, pp. 311–318.
- [19] 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.
- [20] M. Norré, V. Vandeghinste, P. Bouillon, T. François, Extending a text-to-pictograph system to french and to arasaac, in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, INCOMA Ltd., 2021, pp. 1050–1059.
- [21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, *OpenAI Blog* 1 (2019) 9.
- [22] M. Popović, chrF: character n-gram f-score for automatic mt evaluation, in: *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.
- [23] M. Snover, B. Dor, R. Schwartz, L. Micciulla, J. Makhoul, 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*, 2006, pp. 223–231.
- [24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
- [25] A. Lavie, G. Hanneman, S. Agrawal, D. Kanojia, C.-K. Lo, V. Zouhar, F. Blain, C. Zerva, E. Avramidis, S. Deoghare, A. Sindhuja, J. Wang, D. I. Adelani, B. Thompson, T. Kocmi, M. Freitag, D. Deutsch, Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help, in: *Proceedings of the Tenth Conference on Machine Translation*, Association for Computational Linguistics, Suzhou, China, 2025, pp. 436–483. URL: <https://aclanthology.org/2025.wmt-1.24>. doi:10.18653/v1/2025.wmt-1.24.

A. LLM Evaluation Prompt

The following system prompt (Figure 3) was used for LLM-based evaluation of pictogram label translation using Qwen2.5-72B-Instruct.

You are an expert evaluator for a Text-to-Pictogram Label Translation system. In this task, a natural language sentence is converted into a sequence of pictogram labels (hyp). A human reference label sequence (tgt) is provided as the ground truth.

Important characteristics of pictogram labels:

- They may use underscores instead of spaces
(bus_station = bus station)
- They may omit function words (articles, prepositions, conjunctions) that have no pictogram representation
- Word order may differ slightly from tgt while preserving the same meaning
- Synonymous labels are acceptable:
village ~ municipality ~ town
locate ~ appear ~ turn up
Great_Britain ~ UK ~ United_Kingdom
- Extra labels in hyp are acceptable if they extend the meaning of tgt
- Missing labels are penalised only when they drop a core concept from tgt

You must evaluate on exactly ONE dimension.

ADEQUACY (1-5)

Does hyp convey the same meaning as tgt, tolerating synonymous label variants?

- 1 = hyp is largely unrelated to tgt in content
- 2 = hyp shares some concepts with tgt but misses most key elements
- 3 = hyp captures the main idea of tgt but has noticeable gaps or wrong labels
- 4 = hyp matches tgt well; only minor synonym differences or small additions
- 5 = hyp is semantically equivalent to tgt (synonyms and label-style variants OK)

After the score, provide a brief reasoning (2-3 sentences) citing specific label examples from tgt and hyp to justify your rating.

Respond ONLY with a valid JSON object - no markdown, no extra text:

```
{  
  "adequacy": <integer 1-5>,  
  "reasoning": "<string: 2-3 sentence explanation>"  
}
```

Figure 3: System prompt used for LLM-as-Judge evaluation with Qwen2.5-72B-Instruct.

B. Examples from Qualitative Analysis

Table 5 presents selected examples from the official test set, translated by our best-performing system and evaluated using different metrics. In Example 1, the hypothesis uses *municipality*, whereas the reference uses *village*. As expected, BLEU assigns a lower score due to the lexical mismatch, while BERTScore assigns a higher score because the two words are semantically similar. Likewise, Qwen considers the reference and hypothesis semantically equivalent and assigns the highest score (5). chrF also performs better than BLEU by awarding partial credit. Interestingly, *village* and *municipality* are represented by the same pictogram in ARASAAC, making the reference and hypothesis identical from a visual perspective. Example 2 shows a similar case where semantic similarity results in high scores from embedding-based and LLM-based metrics. However, the label *humanitarian* does not exist in the ARASAAC pictogram set, a discrepancy that can only be identified using multimodal information. This example illustrates that semantic similarity alone may lead to misleading evaluations.

Examples 3 and 4 demonstrate cases where embedding-based metrics assign high scores despite important differences between the reference and hypothesis.

Example 5 further motivates the need for a visual evaluation dimension. Although *metropolis* and *city* are represented by the same pictogram in ARASAAC, text-based metrics penalise this lexical variation.

Finally, Example 6 shows that lexical metrics penalise acceptable variations between the reference and hypothesis, whereas metrics incorporating semantic information produce more appropriate evaluations.

N	ref	hyp	BLEU	TER	BERT_P	BERT_R	chrF6	chrF4	Qwen
1	the village of metropolis place above Madagascar	the municipality of metropolis place above the Madagascar	38.26	28.57	0.91	0.93	71.38	76.38	5
<i>Qwen Reasoning:</i> The hyp uses 'municipality' as a synonym for 'village', and both sequences correctly identify 'metropolis place' and its location 'above the Madagascar'. The meaning is semantically equivalent.									
2	she work in the volunteer and at_the Afghanistan	she work in the humanitarian and the Afghanistan	33.2771	25.0	0.92	0.88	58.77	64.88	4
<i>Qwen Reasoning:</i> The hyp captures the main idea of tgt, with 'humanitarian' being a reasonable synonym for 'volunteer'. The only minor difference is the omission of 'at_the', which does not change the meaning.									
3	past half_body take_part himself to the indige-nous_person_of_Asia	past he himself take_part to the rook by Asia	11.43	71.44	0.83	0.78	37.18	46.81	3
<i>Reasoning:</i> The hyp captures some key elements like 'past', 'himself', and 'take_part', but replaces 'indige-nous_person_of_Asia' with 'rook by Asia', which changes the meaning.									
4	the fabric_tunnel upgrade under the parliamen-tary_group	the tunnel pro-mote under the monument	9.91	50.0	0.80	0.75	30.17	38.1152	2
<i>Reasoning:</i> The hyp shares partial concepts like 'tunnel' and 'under', but misses key elements such as 'fabric_tunnel' and 'parliamentary_group'. 'promote' does not correctly reflect 'upgrade'.									
5	he burial at_the city	he burial at_the metropolis	75.98	25.0	0.94	0.9553	70.28	73.11	4
<i>Reasoning:</i> The hyp preserves the meaning of tgt; 'metropolis' is a valid synonym for 'city' with no significant meaning change.									
6	she have_got right order in_order_to the Australia	she possess well order for the Aus-tralia	6.63	42.86	0.88	0.79	34.87	41.66	4
<i>Reasoning:</i> The hyp maintains the core meaning; 'possess' corresponds to 'have got' and 'well' loosely maps to 'right'. Minor structural differences do not affect meaning.									

Table 5

Examples from the qualitative analysis showing automatic metric scores, Qwen judgments, and the corresponding Qwen reasoning. BERT_P and BERT_R denote the BERTScore precision and recall, respectively. chrF6 and chrF4 denote chrF computed with character n-grams of 6 and 4, respectively. Qwen refers to the Qwen2.5-72B-Instruct model used as the LLM-based evaluator.