Overview of the 2026 ImageCLEFtoPicto Task: Investigating the Generation of Pictogram Sequences from an English Text Dataset

Notebook for the ImageCLEF Lab at CLEF 2026

Maja Hjuler[1,2,*], Diandra Fabre[1], Claire Lemaire[3,1], Benjamin Lecouteux[1] and Didier Schwab[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2]School of Computer Science, Queensland University of Technology, Brisbane, Australia
[3]Univ. Toulouse, LAIRDIL, 31000 Toulouse, France

Abstract

The automatic generation of a pictogram sequence from a text input is a novel challenge in the NLP community. It has the potential to support communication for people with complex communication needs, relying on Augmentative and Alternative Communication. This paper presents an overview of the third edition of the ImageCLEFtoPicto task that includes two sub-tasks: Text-to-Picto, whose goal is to produce a comprehensive sequence of pictogram terms given a text input, and Next-Pictogram Prediction, which focuses on predicting the most probable next pictogram in a given sequence. The main change compared to previous editions is a shift in source language from French to English. As no aligned English text–pictogram resource or system targeting ARASAAC previously existed, this edition establishes, to our knowledge, the first community-level benchmark for English Text-to-Picto translation. The Speech-to-Picto sub-task offered in the second edition is not part of this edition. Text-to-Picto attracted highly competitive submissions, with the top systems clustered within fractions of a point (best run: 52.55 sacreBLEU, 71.69 METEOR, 31.48 PictoER). This suggests the task is approaching a performance ceiling under current modelling approaches. The newly introduced Next-Pictogram Prediction sub-task proved considerably harder, with the best system reaching 8.91% Top-1 accuracy, reflecting the scarcity of training data relative to mainstream language modeling.

Keywords

ImageCLEF, Pictograms, Automatic Translation, Augmentative and Alternative Communication, Multimodal Data, Text-to-Picto, Next-Pictogram Prediction

1. Introduction

The ImageCLEFtoPicto task [1] is part of the ImageCLEF lab[1] initiative, which aims to advance the evaluation of technologies across a range of tasks (annotation, generation, classification). It offers reusable benchmarking resources based on a large collection of multimodal data, supporting evaluations in monolingual, cross-language, and language-independent contexts. The toPicto task is the third edition of the task, with the goal of generating a pictogram translation from a text input. This natural language processing (NLP) challenge introduces a novel type of multimodal data for training machine learning models. The main change compared to previous editions is a shift in source language from French to English. While text-to-pictograph systems have previously been developed for other languages [2] and adapted to French and ARASAAC (Aragonese Center of Augmentative and Alternative Communication) [3], no aligned English text–pictogram resource or automatic English Text-to-Picto system targeting ARASAAC existed at the outset of this work. To our knowledge, ToPicto 2026 constitutes the first community-level benchmark for English Text-to-Picto translation. Participants are asked to generate a sequence of pictogram terms from English text, adapting their systems to the linguistic,

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
maja-jonck.hjuler@univ-grenoble-alpes.fr (M. Hjuler); diandra.fabre@univ-grenoble-alpes.fr (D. Fabre); claire.lemaire@univ-grenoble-alpes.fr (C. Lemaire); benjamin.lecouteux@univ-grenoble-alpes.fr (B. Lecouteux); didier.schwab@univ-grenoble-alpes.fr (D. Schwab)

[1]https://www.imageclef.org/

syntactic, and semantic specificities of English. In addition to the Text-to-Picto translation sub-task, this edition introduces a new Next-Pictogram Prediction sub-task, which focuses on predicting the most probable next pictogram in a given sequence. The Speech-to-Picto sub-task offered in the second edition is not part of this edition.

Pictogram-based translation is an emerging NLP task that has recently drawn research interest for its potential to help people with complex communication needs convey their messages [4]. For many such individuals, a communication disorder disrupts the ordinary production or comprehension of spoken, written, or signed language [5], whether as a congenital condition present from birth or as an acquired condition arising after language has already developed. In both cases, Augmentative and Alternative Communication (AAC) provide capacity techniques that supplement or replace spoken and written language [6]. AAC is correspondingly used both to build communicative competence for the first time and to recover it after its loss. The pictogram is a central element, a simplified iconic representation of a concept (single word, multi-word expression, named entity, or entire sentence) designed to resemble reality. Beyond conveying basic needs, AAC increasingly aims to support the fuller range of communicative purposes that matter in everyday life, including storytelling [7], sharing experiences, and maintaining relationships [8]. These dimensions bear directly on users' social participation and well-being. Despite these strengths, numerous environmental and financial barriers remain, such as the limited time available for caregivers to learn and teach the use of an AAC tool. Tools that automatically translate speech or text into pictograms can support interactions between AAC users and people in their community with limited or no AAC experience. This edition of the ToPicto task seeks to address this need by fostering the development of methods that generate accurate pictogram sequences from English text and that predict plausible next pictograms, thereby helping reduce communication barriers.

This paper presents an overview of the ImageCLEFtoPicto task, and is organized as follows. We first introduce the two sub-tasks in Section 2, Text-to-Picto and Next-Pictogram Prediction. In Section 3, we explain the creation of the dataset. The evaluation methodology is presented in Section 4. Section 5 describes the participant results with a discussion, before concluding in Section 6 with insights into future directions. The datasets and the scripts to evaluate the submissions are available in our official Hugging Face repository[2].

2. Task descriptions

2.1. Sub-task 1: Text-to-Pictogram Translation

The Text-to-Picto sub-task focuses on the automatic generation of a corresponding sequence of pictogram terms from an English text. This challenge can be viewed as a translation problem, where the source language is English, and the target language is a sequence of English terms. Each of those terms is paired to an ARASAAC pictogram, as illustrated in Figure 1. The provided translation has to follow the specifications regarding a translation in pictograms, understandable by AAC users. ARASAAC is an online resource of over 25,000 pictograms under a Creative Commons license (BY-NC-SA), and funded by the Department of Culture, Sports, and Education of the Government of Aragon (Spain)[3].

2.2. Sub-task 2: Next-Pictogram Prediction

The Next-Pictogram Prediction sub-task focuses on predicting the most probable next pictogram in a given sequence. Given a (possibly partial) sequence of pictogram terms, participants are asked to predict the pictogram term that most likely follows. This setting reflects assistive scenarios in which an AAC interface suggests the next pictogram to a user as they compose a message, thereby reducing the effort required to build a sequence. This sub-task is inspired by the work of Pereira et al. [9]. Concretely, given a pictogram sequence with its final token withheld, participants must predict the missing final

[2] https://huggingface.co/ToPicto
[3] https://arasaac.org/

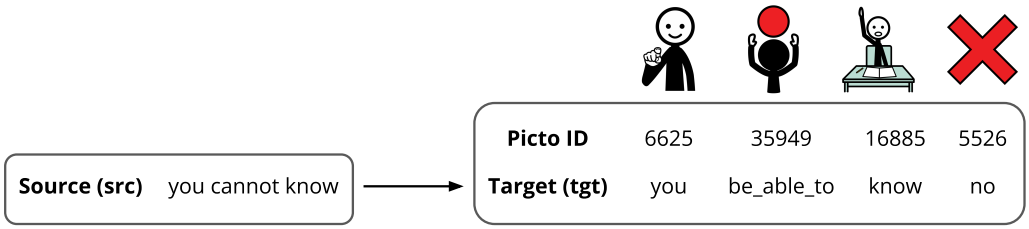

Figure 1: Illustration of the Text-to-Picto sub-task. The input is an English text, and the output is a corresponding sequence of English terms, each of which is paired to a unique ARASAAC pictogram.

Table 1
Format of the Text-to-Picto data.

Tag	Definition	Example
id:	unique identifier of each utterance	cefc-tcof-Acc_del_07-1
src:	source of the utterance: text from oral transcription	you cannot know
tgt:	target of the utterance: sequence of pictogram terms	you be_able_to know no
pictos:	a list of pictogram identifiers linked to each pictogram term (the size is the same as the target output)	[6625, 35949, 16885, 5526]

token. Systems are evaluated under a top-N accuracy setting, so a ranked list of candidate pictogram terms is expected for each instance.

3. Dataset Description

3.1. Source Corpora

The data for the task is derived from the ProPicto-Orféo corpus [10], built from the Corpus d'Étude pour le Français Contemporain (CEFC) [11]. It comprises spontaneous French speech drawn from a range of situations (dialogues, meetings, and other interactional settings), for which a corresponding sequence of pictogram terms, each linked to an ARASAAC pictogram, was produced. This type of spontaneous, conversational text mimics the interactions between individuals who rely on pictograms for communication support and their conversation partners. For this edition, the English text–pictogram parallel data was constructed by translating a subset of the existing ProPicto-Orféo corpus into English, yielding approximately 67,000 sentences. The French source transcripts were translated using an instruction-tuned large language model explicitly prompted to preserve disfluencies and fillers, so that the translated source retains the spontaneous-speech character of the original and remains consistent with its pictogram target. Because ARASAAC pictograms are treated as language-independent, the pictogram sequences themselves are retained from the French data; only the source transcripts are translated, and the target pictogram terms are mapped from French to English via the ARASAAC API[4]. As in previous editions, each source utterance is paired with a target sequence of pictogram terms, each of which is aligned to a unique ARASAAC pictogram identifier.

3.2. Dataset Format

The data for the Text-to-Picto sub-task is contained in a JSON file, which includes the information presented in Table 1 (for training and validation data; for the test set, only the `id` and `src` are provided). The expected output is a JSON file giving, for each `id`, a hypothesis (`hyp`) corresponding to the predicted sequence of pictogram terms. For the Next-Pictogram Prediction sub-task, the data is derived from the

[4]https://arasaac.org/developers/api

Table 2
Statistics of the Text-to-Picto train, validation, and test splits, covering the source texts (`src`) and the target pictogram sequences (`tgt`).

Metric	Train	Validation	Test
Number of utterances	59,278	4,397	4,305
Min length (words in `src`)	1	2	2
Min length (tokens in `tgt`)	3	3	3
Max length (words in `src`)	20	23	19
Max length (tokens in `tgt`)	8	8	8
Average length (words in `src`)	8.03	7.94	7.91
Average length (tokens in `tgt`)	6.40	6.34	6.33
Unique words in `src`	46,156	8,625	8,430
Unique tokens in `tgt`	6,721	2,969	2,912
Unique pictograms in `picto_id`	5,429	2,570	2,500

same underlying corpus as Text-to-Picto. At test time, the model is given all but the last token of a pictogram sequence and must predict the missing final token; the expected output is a ranked list of candidate pictogram terms for each instance. To prevent participants from recovering the withheld final token by matching the input against the source corpus, the `id` field of the Next-Pictogram Prediction data serves as a surrogate identifier rather than the original corpus identifier.

3.3. Dataset Statistics

The statistics of the Text-to-Picto train, validation, and test splits are described in Table 2. The table reports information about the source texts (`src`) and the target pictogram sequences (`tgt`).

The full Text-to-Picto corpus comprises approximately 68,000 sentences, split into 59,278 training, 4,397 validation, and 4,305 test utterances. Two structural properties distinguish the task from conventional machine translation. First, the target pictogram sequences are consistently shorter and more constrained than their source sentences: in the training split, source sentences range from 1 to 20 words and average 8.03 words, while the target sequences range from only 3 to 8 tokens and average 6.40 tokens. This compression is stable across splits, with average target lengths of 6.34 and 6.33 tokens on validation and test, respectively. Second, the source and target operate over vocabularies of very different sizes: the 46,156 unique source words in the training split are reduced to only 6,721 distinct target terms, themselves mapping to 5,429 unique ARASAAC pictograms, and the same asymmetry holds on validation and test. Together, these properties indicate that the task is one of semantic compression into a fixed pictographic inventory rather than word-for-word translation. The Next-Pictogram Prediction sub-task reuses the same underlying training data (approximately 68,000 sentences); its test file contains 1,986 instances, each obtained by withholding the final token from a pictogram sequence.

4. Evaluation Methodology

The evaluation of the Text-to-Picto sub-task is conducted using sacreBLEU [12], METEOR [13], and the Picto-term Error Rate (PictoER) [14, 15], retained from the previous edition for comparability across editions. For all three metrics, the evaluation compares the hypothesis (hyp) provided by the participant with the target (tgt), i.e., the sequence of pictogram terms. All three are computed over the full pictogram-term sequences, treating each space-separated pictogram term (including compound, underscore-joined identifiers such as `to_be_able`) as a single token. We detail each metric and its computation below. The Next-Pictogram Prediction sub-task is evaluated using top-N accuracy, reporting Top-1, Top-3, Top-5, and Top-10 accuracy, complemented by a semantic similarity score that credits predictions that are semantically close to the withheld pictogram even when they do not match exactly. Top-1 accuracy is the primary ranking metric for this sub-task.

A known limitation of all three Text-to-Picto metrics is that they operate at the level of pictogram terms and therefore penalise lexical substitutions even when these are communicatively equivalent. Because several English words may legitimately map to the same ARASAAC pictogram (for example, *tired*, *exhausted*, and *weary* all correspond to a single *fatigue* pictogram), a system can produce a valid translation that is nonetheless scored as an error against the reference term. As the reference targets are derived automatically rather than from independent human annotation, these metrics measure agreement with the dataset's pictogram conventions rather than communicative adequacy as judged by AAC users. The semantic-similarity score introduced for Next-Pictogram Prediction partially mitigates this for that sub-task, and richer adequacy-oriented evaluation is a direction for future editions.

SacreBLEU measures the number of common n-grams (the percentage of overlap) between the translation hypothesis (hyp) and the reference translation (tgt). In comparison with BLEU [16], sacreBLEU takes into account different tokenizations of the same term. The score corresponds to:

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

with:

- BP (Brevity Penalty) is a length penalty that discourages translations that are too short,
- p_n is the modified n-gram precision,
- w_n is the weight associated with n-grams of length n,
- N is the maximum size of n-grams.

The modified n-gram precision is the division of the sum of correct n-grams over their total number in the corpus, with:

$$p_n = \frac{\sum_{\text{n-gram}} \min(\text{Count}^{y}_{\text{n-gram}}, \text{Count}^{x}_{\text{n-gram}})}{\sum_{\text{n-gram}} \text{Count}^{y}_{\text{n-gram}}} \tag{2}$$

where:

- $\text{Count}^{y}_{\text{n-gram}}$ is the number of occurrences of a given n-gram in the translation hypothesis,
- $\text{Count}^{x}_{\text{n-gram}}$ is the number of occurrences of the same n-gram in the reference translation.

The length penalty BP, with the total number of words r in the reference and c the number of words in the translation hypothesis, corresponds to:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases} \tag{3}$$

The score therefore fluctuates between 0 and 1, the highest corresponding to an equivalent translation between hypothesis and reference. In practice, we report sacreBLEU on the 0–100 scale using the reference implementation provided by the Hugging Face `evaluate` library.

METEOR performs an alignment between the translation hypothesis and the reference translations, going beyond simple word matching. It considers not only direct matches but also those based on synonyms, surface form (words that share the same root), and radical form. The evaluation provides more granularity in assessing the performance of an overall system, as it captures additional semantic information that is not encoded within the sacreBLEU score.

METEOR combines the precision and recall of unigrams, as well as a measure to evaluate the word order in the translation compared to the reference. Specifically, the metric combines precision P and recall R of unigrams through a harmonic mean:

$$F_{\text{mean}} = \frac{10 \times P \times R}{R + 9P} \tag{4}$$

A penalty measure for a given alignment is added, with N_{chunks}, the number of contiguous aligned word segments and $N_{unigrams}$ the total number of aligned unigrams:

$$Penalty = 0.5 \times \left(\frac{N_{\text{chunks}}}{N_{\text{unigrams}}} \right)^3 \tag{5}$$

METEOR penalizes alignments where the word order differs or when the aligned words are very far apart in the sentence. This score favors translations that approximate the word order of the reference sentence. The final score is given by:

$$\text{METEOR} = F_{mean} \times (1 - Penalty) \tag{6}$$

The score is a measure between 0 and 1; the higher, the better.

PictoER is a metric derived from the Word Error Rate (WER). Rather than scoring at the level of natural-language words, it is computed over the sequence of pictogram terms, so that each pictogram term (including compound, underscore-joined ARASAAC identifiers) is treated as a single token. The score is defined as follows:

$$\text{PictoER} = \frac{S + D + I}{N} \tag{7}$$

with S the substitutions, I the insertions, D the deletions and N the total number of reference tokens. Concretely, PictoER is obtained by applying the standard WER implementation of the Hugging Face `evaluate` library to the predicted and reference pictogram-term sequences. A lower PictoER indicates better model performance.

Top-N accuracy is used to evaluate the Next-Pictogram Prediction sub-task. Each test instance withholds the final pictogram term of a sequence, and systems return a ranked list of candidate terms. For a given N, an instance is counted as correct if the withheld ground-truth term appears among the system's top-N candidates, and Top-N accuracy is the fraction of evaluated instances for which this holds. We report Top-1, Top-3, Top-5, and Top-10 accuracy, with Top-1 accuracy serving as the primary ranking metric. Before comparison, both the ground-truth term and the candidate terms are normalised by lowercasing and stripping surrounding whitespace, so that matching is case- and whitespace-insensitive. Instances for which a system provides no prediction are excluded from its denominator rather than counted as errors.

Semantic similarity complements top-N accuracy by giving partial credit to predictions that are close in meaning to the withheld pictogram term, even when they do not match exactly. For each instance, we embed the system's top-1 candidate and the ground-truth term using the `all-MiniLM-L6-v2`[5] sentence-transformer model and compute the cosine similarity between the two embeddings. The reported score is the mean cosine similarity over all evaluated instances. As with top-N accuracy, terms are lowercased and stripped before embedding, and instances without a prediction are excluded. Higher values indicate predictions that are, on average, more semantically aligned with the reference.

5. Participant Results and Discussion

The registration and submission of participants' runs were handled by the challenge platform AI4MediaBench[6], developed by AIMultimediaLab[7]. Ten teams submitted runs to the Text-to-Picto sub-task, and eight teams submitted runs to the Next-Pictogram Prediction sub-task. The results, reporting the best run per team, are presented in Table 3 for Text-to-Picto and Table 4 for Next-Pictogram Prediction. All scores are reported as percentages on a 0–100 scale.

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[6] https://ai4media-bench.aimultimedialab.ro/
[7] https://aimultimedialab.ro/

Table 3
ToPicto 2026 Text-to-Picto results, best run per team, ordered by PictoER (lower is better). *Co-authors of the same working note submission.

Rank	Team	sacreBLEU ↑	METEOR ↑	PictoER ↓
1	psp075	52.55	71.69	31.48
2	jlcuad	52.00	71.02	31.99
3	gokul_n_v	52.21	71.04	32.12
4	kaiyili*	51.34	70.40	32.55
5	fuchuanye*	51.34	70.40	32.55
6	sami_haq	51.33	70.36	32.75
7	hiwiy	50.38	69.67	33.33
8	melody	48.49	69.98	35.62
9	lekshmivitscope	28.76	63.92	38.84
10	varghesekjames	31.23	61.18	53.77
11	linxiaocan	22.73	48.59	57.79

Table 4
ToPicto 2026 Next-Pictogram Prediction results, best run per team, ordered by Top-1 accuracy (higher is better). *Co-authors of the same working note submission.

Rank	Team	Top-1 ↑	Top-3 ↑	Top-5 ↑	Top-10 ↑	Sem. Sim. ↑
1	melody	8.91	14.65	18.08	22.21	34.38
2	fuchuanye*	7.85	12.84	14.80	19.34	32.36
3	hiwiy	7.80	12.39	14.15	16.97	33.37
4	kaiyili*	7.25	12.49	15.21	19.74	31.94
5	linxiaocan	6.50	12.64	16.11	21.80	32.50
6	tce_chidambaram	5.44	10.42	14.30	19.34	32.41
7	saisaranya	5.29	10.27	13.39	16.41	32.46
8	sami_haq	3.78	9.47	12.89	19.89	27.02

For Text-to-Picto, the top-ranking teams achieved remarkably close performances, separated by only fractions of a point. The best run, submitted by team `psp075`, reached 52.55 sacreBLEU, 71.69 METEOR, and 31.48 PictoER, narrowly ahead of `jlcuad` and `gokul_n_v`. The considerably lower scores in Next-Pictogram Prediction reflect the intrinsic difficulty of the task, owing in part to the scarcity of training data relative to mainstream language modeling tasks. The best system, submitted by team `melody`, reached 8.91% Top-1 accuracy.

Most participating systems framed Text-to-Picto as a sequence-to-sequence translation problem. They fine-tuned pretrained encoder-decoder transformers such as BART and FLAN-T5 using subword tokenization and beam search decoding. A recurring theme among the stronger submissions was careful handling of the domain-specific compound pictogram tokens, which the default subword tokenizers tend to fragment, as well as attention to data integrity through deduplication and the removal of train–validation overlaps. The narrow spread at the top of the Text-to-Picto leaderboard suggests that the task is approaching a performance ceiling under current modeling approaches. In contrast, the low absolute scores in Next-Pictogram Prediction indicate substantial headroom and the need for approaches better suited to data scarcity.

5.1. Working Notes Summary: Text-to-picto translation

Every submission to the Text-to-Picto subtask framed it as a supervised sequence-to-sequence translation problem and fine-tuned a pre-trained encoder-decoder transformer, with the T5 family clearly dominant. At the top of the official leaderboard (Table 3), Sudharshan PS [17] compared FLAN-T5-base against a BART-base baseline, with FLAN-T5 the stronger of the two. Additionally, they include an extensive exploratory data analysis and a comparison of SentencePiece and BPE tokenizers. HULAT-UC3M by

Guzman-Colomina et al. [18] fine-tuned T5-base after a systematic sweep over base model, task prefix, seed, learning rate, and effective batch size, selecting the final checkpoint on validation generation metrics rather than validation loss. Their prefixed T5-base with greedy decoding ranked second on the official leaderboard. The closely competitive third-ranked system, from Gokul et al. [19], instead used `facebook/bart-base`, chosen for its prefix-free design, and paired it with explicit dataset decontamination, vocabulary augmentation for fragmented pictogram tokens, and a decoding sweep. The only departure from the text-only paradigm came from Ul Haq and Castilho [20], who augmented a T5-small baseline with a visual grounding objective that aligned the decoder with frozen CLIP image embeddings.

A recurring theme was how to handle the closed ARASAAC vocabulary and the mismatch between how compound identifiers[8] should be treated (as atomic) and how subword tokenizers actually slice them (as fragmented). The stronger teams addressed this through vocabulary augmentation: Gokul et al. [19] added frequently fragmented tokens as atomic entries and resized the embeddings, while Adithiyaa D. et al. [21] injected 2,041 compound pictogram tokens directly into the vocabulary. A second strand enforced lexical validity at inference through constrained decoding, either as a hard prefix-tree (Trie) mask over beam search (TokaTron, James et al. [22]) or a soft out-of-vocabulary penalty (Adithiyaa D. et al. [21]). TokaTron found this fix to be strongly tokenizer-dependent and can collapse performance when the identifier segmentation does not match the model's tokenizer. Departing from the text-only paradigm, Ul Haq and Castilho [20] introduced visual grounding, augmenting a T5 baseline with frozen CLIP pictogram-image embeddings aligned through an auxiliary loss, and complemented it with a richer multi-metric evaluation spanning n-gram-based, character-level, learned, and LLM-based measures beyond the official suite. Overall, the subtask drew a methodologically coherent field. The strongest systems cluster tightly around BLEU 52 and METEOR 0.71 despite differing in base model, decoding strategy, and vocabulary handling.

5.2. Working Notes Summary: Next-picto prediction

The Next-Pictogram Prediction subtask, newly introduced this edition, attracted a smaller and more methodologically varied field than Text-to-Picto. Two teams contributed deliberately simple statistical baselines built on the small, closed pictogram vocabulary. Jishnu et al. [23] used a trigram backoff n-gram model with Laplace smoothing, falling back to bigram and unigram statistics. They augment the model with source-word boosting, promoting any candidate pictogram whose name matches a word appearing in the source sentence. Li et al. [24] predicted the next pictogram by computing unigram, bigram, trigram, and overall-frequency probabilities and mixing them in fixed proportions into one score. Huang et al. [25] entered both subtasks and contributed the broadest architecture comparison in the task. For next-pictogram prediction, they evaluated five neural paradigms: text-only Transformer decoding, pictogram-ID modelling, a dual-modal token-ID fusion Transformer at several depths, GPT-2-medium fine-tuning, and a SASRec-style sequential recommender. They find that fusing lexical keyword tokens with discrete pictogram-ID histories outperformed using either alone, and that deeper models performed consistently better. Their best system, a 24-layer dual-modal model, ranked third officially.

Across these approaches, absolute scores remained low (best run: 8.91% Top-1), and the gap between the simple statistical baselines and the strongest neural systems was narrow. This points to data scarcity, rather than model capacity, as the main limiting factor, with room for improvement in future editions through larger training sets or joint modelling of source text and pictogram history.

6. Conclusion and Perspectives

The third edition of the ImageCLEFtoPicto task continued the previous editions' challenge of generating pictogram sequences for Augmentative and Alternative Communication. The main change this year is

[8] Compound identifiers are ARASAAC's underscore-joined pictogram labels, e.g., `to_be_able` and `brushing_teeth`.

a shift in source language from French to English, encouraging participants to adapt their systems to the specificities of English. No aligned English text–pictogram resource or system targeting ARASAAC previously existed. This edition therefore establishes, to our knowledge, the first community-level benchmark for English Text-to-Picto translation, along with a reusable English text–pictogram dataset. The edition included two sub-tasks: Text-to-Picto and the newly introduced Next-Pictogram Prediction; the Speech-to-Picto sub-task offered in the previous edition was not part of this edition. The Text-to-Picto sub-task drew ten teams whose best systems clustered tightly at the top of the leaderboard (best run: 52.55 sacreBLEU, 71.69 METEOR, 31.48 PictoER), most of them fine-tuning pretrained encoder-decoder transformers and paying particular attention to compound pictogram tokens and data integrity. The newly introduced Next-Pictogram Prediction sub-task attracted eight teams and proved markedly harder, with the best system reaching only 8.91% Top-1 accuracy, underlining the challenge posed by the limited amount of training data. In the future, we plan to explore new directions, such as extending the task to additional languages, reintroducing the speech modality, enlarging the training data for next-pictogram prediction, and developing richer evaluation of the adequacy and usability of generated pictogram sequences for AAC users.

Acknowledgments

Co-funded by the European Union under the Marie Skłodowska-Curie Grant Agreement No 101081465 (AUFRANDE). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency, which cannot be held responsible for them. This project was funded by the Agence Nationale de la Recherche (ANR) through the project PANTAGRUEL (ANR-23-IAS1-0001). This work is also carried out as part of the AugmentIA Chair, hosted by the Grenoble INP Foundation with sponsorship from the Artelia Group. The chair also receives support from the French government, managed by the National Research Agency (ANR), under the France 2030 program with reference ANR-23-IACL-0006 (MIAI Cluster). The pictographic symbols used are the property of the Government of Aragón and have been created by Sergio Palao for ARASAAC (http://www.arasaac.org), which distributes them under Creative Commons License BY-NC-SA.

Declaration on Generative AI

Claude (Anthropic) and ChatGPT (OpenAI) were used to assist in editing and polishing the manuscript. All authors reviewed the full manuscript, take responsibility for its content, and consent to its submission.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, 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.

[2] L. Sevens, V. Vandeghinste, I. Schuurman, F. Van Eynde, Extending a dutch text-to-pictograph converter to English and Spanish, in: Proceedings of SLPAT 2015: 6th Workshop on Speech

and Language Processing for Assistive Technologies, Association for Computational Linguistics, Stroudsburg, PA, USA, 2015.

[3] 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), Online, 2021.

[4] M. Romski, R. A. Sevcik, Augmentative communication and early intervention: Myths and realities, Infants & Young Children 18 (2005) 174–185.

[5] L. Cummings, Communication disorders: A complex population in healthcare, Language and Health 1 (2023) 12–19.

[6] D. R. Beukelman, J. C. Light, Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs, 5 ed., Brookes Publishing Co., Baltimore, MD, 2020.

[7] M. Fontana de Vargas, K. Moffatt, Automated generation of storytelling vocabulary from photographs for use in AAC, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1353–1364. URL: https://aclanthology.org/2021.acl-long.108/. doi:10.18653/v1/2021.acl-long.108.

[8] T. M. Weinberg, K. Kadoma, R. E. Gonzalez Penuela, S. Valencia, T. Roumen, Why so serious? exploring timely humorous comments in AAC through AI-powered interfaces, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3714102. doi:10.1145/3706598.3714102.

[9] J. A. Pereira, D. Macêdo, C. Zanchettin, A. L. I. de Oliveira, R. do Nascimento Fidalgo, PictoBERT: Transformers for next pictogram prediction, Expert Systems with Applications 202 (2022) 117231. URL: https://www.sciencedirect.com/science/article/pii/S095741742200611X. doi:10.1016/j.eswa.2022.117231.

[10] C. Macaire, C. Dion, L. Ormaechea, J. Arrigo, C. Lemaire, E. Esperança-Rodier, B. Lecouteux, D. Schwab, Propicto, 2024. URL: https://hdl.handle.net/11403/propicto/v1.1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[11] ORTOLANG, ORFEO: Corpus d'Étude pour le Français Contemporain (CEFC), https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html, 2026. URL: https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html, documentation portal, accessed: 2026-04-29.

[12] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

[13] 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, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909.

[14] J. Woodard, J. Nelson, An information theoretic measure of speech recognition performance, in: Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA, 1982.

[15] A. C. Morris, V. Maier, P. Green, From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition, in: Interspeech 2004, 2004, pp. 2765–2768. doi:10.21437/Interspeech.2004-668.

[16] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[17] P. S. Sudharshan, Transformer-based sequence-to-sequence modeling for English text-to-pictogram translation, CLEF (2026).
[18] I. Guzman-Colomina, J. L. Lopez-Cuadrado, J. L. Lopez-Hernandez, I. Gonzalez-Carrasco, HULAT-UC3M at ImageCLEFtoPicto 2026: English Text-to-Pictogram Generation with T5, CLEF (2026).
[19] N. V. Gokul, S. Rohith, P. Mirunalini, S. Balakumaran, Fine-tuning BART for spoken utterance to pictogram sequence generation in imageclef 2026, CLEF (2026).
[20] S. U. Haq, S. Castilho, Towards multimodal text-to-pictogram generation and evaluation, CLEF (2026).
[21] D. Adithiyaa, T. L. Narayanan, L. Kalinathan, Vocabulary-constrained seq2seq translation for text-to-pictogram generation: A fine-tuned T5 approach with compound token injection at clef 2026 ToPicto, CLEF (2026).
[22] K. V. James, S. S. Krishna, M. Sujith, P. Balasundaram, TokaTron at Text to Picto 2026: Sequence-Trie Constrained Decoding with Seq2Seq for English-to-Pictogram Translation, CLEF (2026).
[23] P. L. Jishnu, M. Santhosh, B. L. Saisaranya, Working Note for ImageCLEFtoPicto 2026: A Statistical Backoff Model for Next-Pictogram Prediction, CLEF (2026).
[24] K. Li, Z. Han, F. Ye, T5 fine-tuning for text-to-picto and interpolated n-gram modeling for next pictogram prediction, CLEF (2026).
[25] J. Huang, X. Qin, Z. Han, Hiwiy at ImageCLEFtoPicto 2026: Text-to-Picto Translation and Next-Pictogram Prediction, CLEF (2026).