Overview of the 2025 ImageCLEFtoPicto Task – Investigating the Generation of Pictogram Sequences from Text and Speech

Notebook for the ImageCLEF Lab at CLEF 2025

Cécile Macaire[1,*], Diandra Fabre[1], Benjamin Lecouteux[1] and Didier Schwab[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Abstract

The automatic generation of a pictogram sequence from either text or speech input is a novel challenge in the NLP community. It has the ability to enhance communication for people with language impairments, relying on Augmentative and Alternative Communication. This paper presents an overview of the second edition of the ImageCLEFtoPicto task. It includes two sub-tasks, Text-to-Picto, whose goal is to produce a comprehensive sequence of pictogram terms given a text input, whereas Speech-to-Picto starts from the speech modality. Compared to last year's edition, the focus is on developing robust translation models across a variety of acoustic domains (read and spontaneous speech) as well as in different thematic (medical, everyday life situations). This paper details the task with the datasets and the evaluation metrics, followed by an overview of models and runs submitted by the participating teams and their results. The best team achieved 76.98 and 62.87 sacreBLEU points on Text-to-Picto and Speech-to-Picto subtasks respectively, highlighting significant progress in addressing this still-challenging task.

Keywords

ImageCLEF, Pictograms, Automatic Translation, Augmentative and Alternative Communication, Multimodal Data

1. Introduction

The ImageCLEFtoPicto task 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 second edition of the task, with the goal of generating a pictogram translation from two different inputs: text or speech. This natural language processing (NLP) challenge introduces a novel type of multimodal data for training machine learning models. Compared to last year's edition, the dataset has been expanded to include a wider variety of acoustic domains, ranging from read to spontaneous speech and thematic topics, including both medical and everyday-life contexts. Participants were tasked with building models that are robust across these diverse domains.

Pictogram translation is a new NLP task that has recently gained interest in the research community, as it can help individuals with language impairments to accurately convey their messages [1]. A communication disorder refers to a disruption in one of the processes that allow speakers to produce and listeners to understand spoken, written, or signed messages [2]. These disorders can be induced before, during, or after birth (congenital disability) or in adulthood, when language and communication skills are already developed (acquired disability). In such cases, Augmentative and Alternative Communication (AAC) can be used. AAC is a set of strategies and methods to supplement or compensate for oral language, used to effectively convey one's thoughts, needs, and emotions [3]. The pictogram is a central

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

cecile.macaire@univ-grenoble-alpes.fr (C. Macaire); diandra.fabre@univ-grenoble-alpes.fr (D. Fabre); benjamin.lecouteux@univ-grenoble-alpes.fr (B. Lecouteux); didier.schwab@univ-grenoble-alpes.fr (D. Schwab)

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

element, a simplified iconic representation of a concept (a single word, multi-word expression, named entity, or entire sentence) designed to resemble reality. Beyond improving access to language, AAC also contributes to users' overall well-being. In a recent study conducted by the Swedish research firm Augur and commissioned by Tobii Dynavox, "Exploring the benefits of assistive communication"[2], high-tech AAC doubles the quality of life for users, whether on a physical level (communicating health issues), social level (building relationships), or psychological level (expressing their emotions and personality). Despite these strengths, numerous environmental and financial barriers remain, such as the lack of time to understand and teach the use of an AAC tool for the caregivers. To bridge this gap between oralizing individuals with no prior knowledge about pictograms and AAC users, tools that automatically translate speech or text into pictograms are essential. This year's edition of the ToPicto task seeks to address this need by fostering the development of robust methods to generate accurate pictogram sequences, thereby helping reduce communication barriers in multiple domains. The datasets and the scripts to evaluate the submissions are available in our official Hugging Face repository[3].

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 Speech-to-Picto. In Section 3, we explain the creation of the multimodal dataset. The evaluation methodology is presented in Section 4. Section 5 describes the participant results with a discussion, before concluding in Section 6 with some insights into future directions.

2. Task description

The ImageCLEFtoPicto 2025 task consists of two sub-tasks: Text-to-Picto and Speech-to-Picto. Participants were allowed to submit to one or both sub-tasks, with a maximum of 10 submissions in total.

2.1. Sub-task 1: From Text to Pictogram Sequence

The Text-to-Picto sub-task focuses on the automatic generation of a corresponding sequence of pictogram terms from a French text. This challenge can be viewed as a translation problem, where the source language is French, and the target language is a sequence of French pictogram terms, each linked to an ARASAAC pictogram, as illustrated in Figure 1. 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).

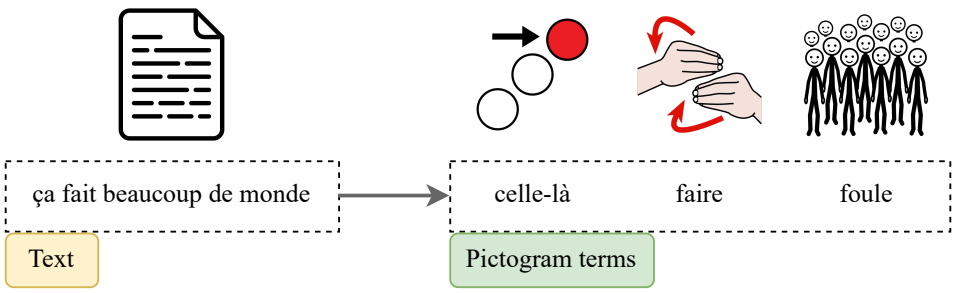

Figure 1: Illustration of the Text-to-Picto sub-task. The input is a French text, and the output is a corresponding sequence of pictogram terms, each of them are linked to a unique ARASAAC pictogram.

[2]https://safecaretechnologies.com/wp-content/uploads/2024/06/AAC_Health_Economic_Study.pdf
[3]https://huggingface.co/ToPicto

2.2. Sub-task 2: From Speech to Pictogram sequence

The Speech-to-Picto sub-task focuses on two modalities: speech and pictograms. Unlike traditional spoken language translation systems that rely on transcribed text, this approach aims to directly map a speech input to pictogram concepts. An example of this task is illustrated in Figure 2.

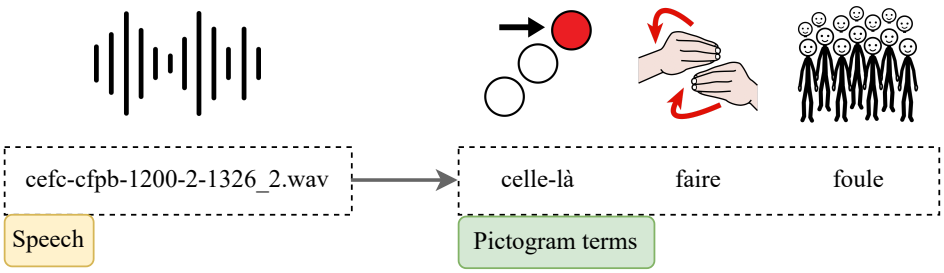

Figure 2: Illustration of the Speech-to-Picto sub-task. The input is a French audio in .wav format, and the output is a corresponding sequence of pictogram terms, each of them are linked to a unique ARASAAC pictogram.

3. Dataset Description

3.1. Source Corpora

The benchmarking data are curated from three corpora containing aligned speech, text, and pictogram sequences: Propicto-commonvoice, Propicto-orféo, and Propicto-eval [4, 5, 6].

Propicto-commonvoice is built from the French portion of the CommonVoice version 15 corpus [7]. It includes 967 hours of recordings of read speech from 17,911 unique speakers. The method described in Macaire et al. [4] was applied to generate a corresponding pictogram translation for each audio segment in the form of a token sequence.

Propicto-orféo is a corpus of spontaneous speech derived from the Corpus d'Étude pour le Français Contemporain (CEFC) [8], including a set of 12 source corpora featuring diverse speech situations (dialogues, meetings, etc.) across various domains. Propicto-orféo comprises 290,036 audio segments totaling 233 hours. The same method [4] was applied to produce pictogram translations from the transcriptions of these audio segments.

Propicto-eval is an evaluation corpus specifically designed to assess the performance of pictogram translation models in a controlled scenario. This multi-speaker read speech corpus incorporates textual data from children's stories, everyday life situations, and sentences taken from the medical domain.

3.2. Dataset Format

The data for both sub-tasks is contained in a JSON file, which includes the following information:

Tag	Definition	Example
id:	unique identifier of each utterance	cefc-tcof-Acc_del_07-1
src:	audio file linked to the ID in .wav format (speech-to-picto) / text from oral transcription (text-to-picto)	cefc-tcof-Acc_del_07-1.wav / tu peux pas savoir
tgt:	target of the utterance — sequence of pictogram terms (tokens)	toi pouvoir savoir non
pictos:	a list of pictogram identifiers linked to each pictogram terms (the size is the same as the target output)	[6625, 35949, 16885, 5526]

The participant's goal is to provide a hypothesis (hyp) equivalent to the target (tgt). To visualize the target as a sequence of pictogram images, we developed an online platform, *Visualize-Pictograms*[4]. If a pictogram token does not exist, the corresponding pictogram image will not be displayed.

3.3. Dataset Statistics

The dataset statistics of the Text-to-Picto and Speech-to-Picto development and test sets are described in Table 1, along with the distribution of data sources. We randomly selected 10,000 utterances from Propicto-orféo and Propicto-commonvoice, as well as medical utterances from the Propicto-eval medical subset.

Table 1
Statistics of the Text-to-Picto and Speech-to-Picto development and test sets. The number of utterances is shown for the train, validation, and test splits, along with the distribution of data sources: Propicto-commonvoice, Propicto-orféo, and Propicto-eval.

		train	valid	test
	Number of utterances	20,177	1,208	2,904
Source distribution:	Propicto-commonvoice	10,000	600	1,475
	Propicto-orféo	10,000	600	1,365
	Propicto-eval	177	8	64

Table 2 gives additional information about the source and target data for the development and test sets. For Text-to-Picto, it includes the minimum and maximum text lengths. For Speech-to-Picto, it specifies the minimum and maximum durations of speech utterances in seconds. The unique tokens represent the number of distinct pictogram tokens in the entire set.

Table 2
Statistics of the source and target data in the development and test sets for the Text-to-Picto and Speech-to-Picto sub-tasks. For source data, the minimum, maximum, and average text lengths (Text-to-Picto) and speech durations in seconds (Speech-to-Picto) are displayed. For target data, it shows corresponding text lengths, as well as the number of unique pictogram tokens.

	Source			Target		
	train	valid	test	train	valid	test
Min length / duration (seconds)	1 / 0.08	1 / 0.22	1 / 0.19	1	1	1
Max length / duration (seconds)	99 / 28.28	50 / 21.97	62 / 21.35	89	48	50
Average length / duration (seconds)	9.8 / 4.00	1.0 / 4.35	10.12 / 4.45	8.5	8.6	8.7
Unique tokens	-	-	-	4,346	1,492	2,308

4. Evaluation Methodology

The evaluation is conducted using sacreBLEU [9], METEOR [10], and the Picto-term Error Rate (PictoER) [11, 12]. 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. We detail each metric and its computation below.

SacreBLEU measures the number of common n-grams (the percentage of overlap) between the translation n hypothesis (hyp) and the reference translation (tgt). In comparison with BLEU [13], sacreBLEU takes into account the meaning of words and different tokenizations of the same term. The score corresponds to:

[4]https://huggingface.co/spaces/ToPicto/Visualize-Pictograms

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

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:

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

The score therefore fluctuates between 0 and 1, the highest corresponding to an equivalent translation between hypothesis and reference.

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 WER. Instead of evaluating the number of errors at the word level, we focus on the number of errors of tokens, each linked to an ARASAAC pictogram. 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 tokens. A lower PictoER indicates better model performance.

5. Participant Results and Discussion

The registration and submission of participants' runs were handled by the challenge platform AI4MediaBench[5], developed by AIMultimediaLab[6]. In this year's edition, 36 teams registered for both sub-tasks. For the Text-to-Picto sub-task, 2 teams submitted their work, resulting in a total of 4 runs, whereas the Speech-to-Picto sub-task received only 2 submissions from a single team. Table 3 presents the overall results of both sub-tasks, sorted by the best sacreBLEU score.

Table 3
Performance of participating teams in the ImageCLEFtoPicto 2025 task. Scores for sacreBLEU, METEOR, and PictoER are reported and ordered by the highest sacreBLEU score.

Sub-Task	Team Name	SacreBLEU↑	METEOR↑	PictoER (%)↓	Rank
Text-to-Picto	majahj	76,98	88,66	13,48	1
	sudharshan07	69,01	85,09	18,56	2
	indira	52,41	74,50	29,23	3
	indira	37,72	64,61	42,70	4
Speech-to-Picto	majahj	62,87	73,41	29,49	1
	majahj	54,71	65,90	40,02	2

The submissions presented in Table 3 are to be divided into two teams: TEAM1, composed of *majahj* and *indira*, and TEAM2, composed of *sudharshan07*. Despite interesting results, TEAM2's paper was rejected due to its lack of references and information on the architecture.

TEAM1 presented a study addressing both the Text-to-Picto and Speech-to-Picto subtasks. The authors built upon one of the submissions from the previous edition of the challenge, presented in Koushik et al. [14]. TEAM1 fine-tuned a T5 encoder-decoder architecture [15] and extended the experiments proposed by Koushik et al. [14] to different T5 sizes (base, small, large) and a larger number of training epochs. For the Text-to-Picto subtask, the T5-large model achieved the best performance (ranking first in Table 3). For the Speech-to-Picto subtask, a cascaded architecture was used, combining a Whisper-based ASR model [16] with the fine-tuned T5-large model. The larger Whisper model (ranking first in Table 3) performs better than the smaller model (ranking second).

Both these results highlight the impact of the size of pre-trained models on downstream tasks.

The authors carefully analyzed the model outputs and identified challenges faced when translating to pictograms, such as handling proper nouns, tense, and numeric expressions. This work opens perspectives for next year's challenge.

6. Conclusion and Perspectives

The second edition of the ImageCLEFtoPicto task continued last year's challenge by offering an expanded version of the dataset, featuring a variety of acoustic scenarios and domains, including medical and everyday-life contexts. The challenge included two subtasks: Text-to-Picto and Speech-to-Picto. Despite

[5] https://ai4media-bench.aimultimedialab.ro/
[6] https://aimultimedialab.ro/

a high number of registrations, actual participation and submissions were limited. Nevertheless, the best submission established a solid baseline and methodology for addressing the task, along with an insightful analysis of the generated pictogram sequences. In the future, we plan to explore new directions, such as providing an English version of the dataset and introducing a next-pictogram prediction subtask.

Acknowledgments

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, led by Didier Schwab and 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), that distributes them under Creative Commons License BY-NC-SA.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o-mini in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

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

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

[3] D. R. Beukelman, P. Mirenda, et al., Augmentative and alternative communication, Paul H. Brookes Baltimore, 2013.

[4] 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: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 839–849. URL: https://aclanthology.org/2024.lrec-main.76/.

[5] C. Macaire, C. Dion, D. Schwab, B. Lecouteux, E. Esperança-Rodier, Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes, in: M. Balaguer, N. Bendahman, L.-M. Ho-dac, J. Mauclair, J. G Moreno, J. Pinquier (Eds.), Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, ATALA and AFPC, Toulouse, France, 2024, pp. 22–35. URL: https://aclanthology.org/2024.jeptalnrecital-taln.2/.

[6] C. Macaire, C. Dion, D. Schwab, B. Lecouteux, E. Esperança-Rodier, Towards speech-to-pictograms translation, in: Interspeech 2024, 2024, pp. 857–861. doi:10.21437/Interspeech.2024-490.

[7] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4218–4222. URL: https://aclanthology.org/2020.lrec-1.520/.

[8] J.-M. Debaisieux, C. Benzitoun, H.-J. Deulofeu, Le projet ORFEO: Un corpus d'études pour le français contemporain, Corpus 15 (2016) 91–114. URL: https://hal.science/hal-01449600. doi:10.4000/corpus.2936.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] A. Koushik, J. Morrison, P. Mirunalini, et al., A transformer based approach for text-to-picto generation (2024).

[15] 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.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.