Parlez-vous Picto? A Transformer-Based Approach for Text-to-Picto and Speech-to-Picto Translation in French

Notebook for the ImageCLEF Lab at CLEF 2025

Maja J. Hjuler[1,2,*,†], Indira Fabre[3,†]

[1]University Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2]School of Computer Science, Queensland University of Technology, Brisbane QLD 4000, Australia
[3]Télécom Paris, 91120 Palaiseau, France

Abstract

This study was conducted in the context of the ToPicto task of ImageCLEF 2025. It investigates the performance of a Transformer-based approach for Text-to-Picto and Speech-to-Picto translation from French language using the pre-trained Google-T5 model fine-tuned on the provided dataset. The T5-large version of the model resulted for the Text-to-Picto task in a score of 93.0, 95.7, and 3.4 for SacreBLEU, METEOR, and PictoER, respectively. To solve the Speech-to-Picto task, this model was combined with a pre-trained ASR model and gave promising results. These findings indicate potential for developing tools to facilitate communication between AAC users and others.

Keywords

Natural Language Processing, Transformer model, Google-T5, French text translation, Pictogram generation

1. Introduction

Augmentative and Alternative Communication (AAC) encompasses methods used to supplement or replace speech and writing, particularly for individuals with speech and language impairments. Existing AAC systems are diverse, with many implementations utilizing pictograms. This visual and symbol-based approach can significantly enhance communication effectiveness and has been shown to be efficient [1]. A key challenge is bridging the gap between AAC users and the broader society. Thus, developing tools that convert speech and text modalities into a sequence of pictograms is essential for facilitating effective communication between these two groups. Recent advances in Transformer models in natural language processing tasks, along with the results from the previous ToPicto Challenge 2024 [2], led us to further explore these architectures for Text-to-Picto and Speech-to-Text-to-Picto translation tasks. This research focuses on using the pre-trained Google T5 model and fine-tuning it with the new corpus provided for the ToPicto task of ImageCLEF 2025 [3, 4].

2. Related Work

Large Language Models (LLMs) have revolutionized various text- and speech-based tasks, including speech recognition, language translation, and augmentative communication systems [5, 6, 7, 8]. In the context of AAC, LLMs based on the Transformer architecture enable more accurate and context-aware language processing. Unlike traditional statistical models, Transformer utilize self-attention mechanisms to capture long-range dependencies, improving speech-to-pictogram translation and next-pictogram prediction. The first study on automatic translation of French speech into a sequence of pictograms was presented by Vaschalde et al. [9]. Their methodology adapts the Text-to-Picto [10] system by integrating four modules: an ASR system, a simplification system, a word sense disambiguation model,

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
maja-jonck.hjuler@univ-grenoble-alpes.fr (M. J. Hjuler); indira.fabre@telecom-paris.fr (I. Fabre)

and a module to display the sequence of pictograms. The automatic translation of speech into pictogram terms (Speech-to-Picto) has the potential to improve communication for individuals with language impairments [11]. For example, this technology can facilitate communication from a non-AAC user to an AAC user, or it can help individuals with speech disabilities learn how to use pictograms for self-expression.

Macaire et al. [11] investigated two approaches for Speech-to-Picto (S2P) translation: (1) the cascade approach combines an Automatic Speech Recognition (ASR) system with a machine translation system, and (2) the end-to-end approach, which tailors a speech translation system to perform direct translation from an audio sequence. Propicto-orféo, described in [12], is used for training after preprocessing by splitting into training, validation, and test sets (80/10/10 split). Propicto-orféo [13] contains 230 hours of French speech resources with speech units aligned to pictograms. The team created another dataset, Propicto-eval, with speech transcriptions from 62 speakers, and used a subset of 100 sentences for the final performance evaluation. Based on BLEU scores [14], the cascade approach outperforms the end-to-end approach. The cascade approach achieves scores of 62.5 and 77.2 on the Propicto-orféo and -eval datasets, respectively, compared to scores of 60.2 and 54.5 for the end-to-end S2P approach.

Previous years' submissions to ImageCLEF have also explored the use of LLMs to solve the Text-to-Picto task. Anand et al. [15] implemented a Transformer model utilizing embeddings from CamemBERT [16], a French BERT model fused with a contrastive learning technique. Elliah et al. [17] finetuned pre-trained translation models (GPT-2 [18] and Helsinki-BERT [19]) for Text-to-Picto conversion, utilizing tokenization and lexical simplification. Similarly, Koushik et al. [20] fine-tuned Google-T5 [21] for the task of translating French text into pictogram sequences. Their proposed model obtained a PictoER score of 13.9, a BLEU score of 74.4, and a METEOR score of 87.1.

3. Dataset

The dataset used in this study is sourced from the CommonVoice v.15 corpus [22] and the Orféo corpus [23]. CommonVoice is a multilingual, publicly available voice dataset recorded by users on the Common Voice platform (http://voice.mozilla.org/). It is intended for speech technology research and development and is based on text from various public domain sources. Only French language data were used from this dataset. Orféo is a corpus consisting of both spoken and written French samples. It contains interactions between adults, adults and children, as well as between children. It has the advantage of being representative of the interactions observed between caregivers and individuals who rely on pictograms due to language impairments. Training, validation, and test splits consist of 20,177, 1,208, and 2,901 utterances, respectively. For the Speech-to-Picto task, a corresponding audio sequence associated with a pictogram sequence is provided (S2P src in Table 1). For the Text-to-Picto translation task, a corresponding sequence of terms associated with a pictogram sequence is provided, derived from the speech transcription (T2P src in Table 1).

Table 1
Description of the Speech-to-Picto and Text-to-Picto dataset, where ID, source (S2P src or T2P src depending on the task), target, and pictograms are available. For the test data, only the ID and source were given.

Tag	Definition	Example
id	unique identifier of each utterance	`common_voice_fr_21455110`
S2P src	audio file linked to the ID in .wav format	`common_voice_fr_21455110.wav`
T2P src	source of the utterance - text from oral transcription	il a découvert deux astéroïdes et une comète
tgt	target of the utterance - sequence of pictogram terms (tokens)	passé il inventer deux pluton et une comète
pictos	a list of pictogram identifiers linked to each pictogram terms (the size is the same as the target output).*	[9839, 6480, 6531, 2628, 10299, 11399, 8474, 2711]

ARASAAC pictograms are used as a reference for pictogram translation. Images can be obtained via the ARASAAC API using https://api.arasaac.org/v1/pictograms/{pictogram_ref_number}.

4. Approach

4.1. Text-to-Picto

This research focuses on addressing the Text-to-Picto task using a text-to-text approach based on the pre-trained Google T5 model, which is fine-tuned on the provided corpus. The output sequence of French terms must correspond to a sequence of French pictogram terms and comply with the specifications of AAC. T5 is an encoder-decoder Transformer available in various sizes, ranging from 60 million to 11 billion parameters [21]. Its ability to handle a wide range of NLP tasks by treating them all as text-to-text problems makes it an attractive choice for Text-to-Picto translation. Unlike other SOTA models, such as BERT [24] and GPT-2 [18], which are primarily designed for specific tasks like language modeling or masked language modeling, T5's unified text-to-text framework allows for greater flexibility and adaptability across tasks.

Table 2
Overview of training details and hardware specifications for each model, including number of Transformer layers, number of model parameters, training time, and the type of GPU used for training.

Model[1]	# trans. layers	# params	# time (h:mm)	GPU
T5-small	12	60M	1:30	4x NVIDIA GTX 1080 Ti (11 GB)
T5-base	24	220M	4:10	4x NVIDIA GTX 1080 Ti (11 GB)
T5-large	48	770M	4:55	4x NVIDIA Quadro RTX 8000 (48 GB)

Fine-tuning is performed using the Seq2SeqTrainer class from the HuggingFace framework[2] [25], with code adapted from Macaire et al.[3] [11]. Table 2 gives an overview of the model training and GPU resources. Other hyperparameters for training include:

- Batch size: 8
- Learning rate: $2 \cdot 10^{-5}$
- Weight decay: 0.01

Both the source data and target pictogram sequences are tokenized using the pre-trained tokenizer corresponding to the size of the T5 model. Padding and truncation are used to ensure text sequence lengths of 128 tokens, and the tokenizer has not been fine-tuned.

4.2. Speech-to-Picto

For the Speech-to-Picto task, the speech is first converted to text using two models of the Whisper family[4] [26] before applying the same Text-to-Picto approach described above. The models used are Whisper-small (244 million parameters) and Whisper-large (1,550 million parameters). We directly use the Whisper models for inference; hence, no model training is involved in this process. The choice to implement a cascade approach (Speech-to-Text followed by Text-to-Picto) rather than directly fine-tuning on the audio was primarily dictated by the limited time available for the project. Furthermore, the work by Macaire et al. [11] suggests that the cascade approach is superior to end-to-end models that directly translate audio into pictogram tokens.

[1] google-t5/t5-small; google-t5/t5-base; google-t5/t5-large
[2] Hugging Face Transformers Seq2SeqTrainer documentation and repository.
[3] macairececile/speech-to-pictograms
[4] openai/whisper-small; openai/whisper-large

5. Evaluation Methodology

The evaluation is conducted using SacreBLEU [14], METEOR [27], and the Picto-term Error Rate (PictoER), derived from the Word Error Rate (WER) [28].

SacreBLEU is a standardized version of the BLEU score, which measures the number of common n-grams between the two sequences. METEOR (Metric for Evaluation of Translation with Explicit ORdering) provides a more nuanced evaluation by incorporating synonymy and stemming and capturing additional semantic information that is not encoded in the BLEU score. PictoER is tailored for evaluating translations involving pictorial terms. Instead of evaluating the number of errors at the word level, it focuses on the number of errors of tokens, each linked to an ARASAAC pictogram.

It is worth noting that this evaluation method does not account for cases where different words or phrases correspond to the same pictogram. For instance, the French words *"épuisé"*, *"exténué"*, and *"fatigué"* all convey similar meanings and are mapped to the same pictogram (displayed in Figure 1). However, under the current evaluation approach, substituting one of these synonyms for another would result in a lower score, despite semantic equivalence. The same limitation applies to numbers: whether expressed as digits (e.g., *"3"*) or in written form (e.g., *"trois"*), they are represented by the same pictogram, yet such variations are still penalized in the scoring.

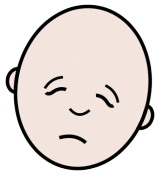


Figure 1: Identical pictogram representing the French words *"épuisé"*, *"exténué"*, and *"fatigué"*.

6. Results and Discussion

The model performance is evaluated using the three different metrics by comparing the predicted pictogram sequence to the target (tgt).

6.1. Text-to-Picto

Table 3
Model performances for the different Google-T5 models trained to perform the Text-to-Picto task. SacreBLEU, METEOR, and PictoER scores are reported for the train, validation, and test sets. Only the models with test scores were submitted to ImageCLEF.

Model	Epoch	SacreBLEU			METEOR			PictoER		
		Train	*Validation*	*Test*	*Train*	*Validation*	*Test*	*Train*	*Validation*	*Test*
T5-small	13	43.2	42.9	37.7	68.3	67.1	64.6	36.9	37.4	42.7
T5-small	20	56.1	53.8	-	76.2	74.4	-	26.8	28.6	-
T5-base	15	55.9	54.5	52.4	76.2	74.6	74.5	26.9	28.6	29.2
T5-base	20	79.0	72.2	-	88.9	85.2	-	11.8	16.6	-
T5-large	19	**93.0**	**80.1**	-	**95.7**	**89.4**	-	3.5	**11.8**	-
T5-large	20	**93.0**	80.0	**77.0**	**95.7**	89.3	**88.7**	**3.4**	**11.8**	**13.5**

The results obtained by fine-tuning T5-base for 15 epochs are lower than those reported by Koushik et al. [20], who achieved scores of 13.9, 74.4, and 87.1 for PictoER, BLEU, and METEOR, respectively, after only 6 epochs. Comparable results are achieved after additional training up to 20 epochs. This discrepancy may be attributed to differences in the data used. Using the T5-large version of the model

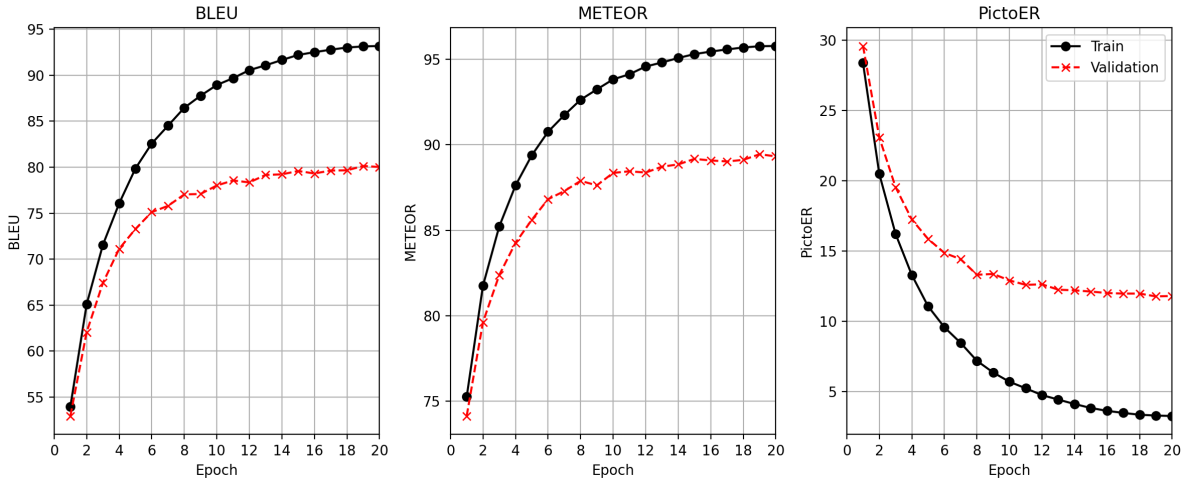


Figure 2: SacreBLEU, METEOR, and PictoER scores for T5-large finetuning for 20 epochs.

significantly improved performance, yielding scores of 93.0, 95.7, and 3.4 for SacreBLEU, METEOR, and PictoER, respectively. However, the scores obtained on the validation and test sets are substantially lower, indicating limited generalization to unseen data. The loss curves presented in Appendix A.3 show a slight increase in validation loss after 10 epochs, a trend that continues until 20 epochs, potentially indicating overfitting. Figure 2 visualizes the performance metrics evaluated on the training and validation sets for each epoch during T5-large fine-tuning. Performance appears to saturate around 15 epochs, with little improvement in evaluation metrics thereafter. A comparison of checkpoints at epochs 19 and 20 of the same training run (Table 2) supports this observation; however, the difference is insignificant compared to the variation between different training runs. An analysis of uncertainty in performance metrics, such as by averaging over training runs, would be necessary to confirm this observation.

Some representative examples are selected for qualitative analysis to highlight behavioral differences between model sizes and inherent challenges associated with text-to-pictogram translation for this particular dataset. The pictogram sequences are generated from predicted tokens using the Hugging Face platform[5]. Extensive analysis can be found in the Appendix B.

Key improvements observed with the T5-large model, compared to smaller versions, include a reduced tendency to generate words that do not correspond to any existing pictogram. Both T5-base and T5-large demonstrate enhanced ability to correctly translate past tense and proper nouns of names and places, which are often translated by a generic pictogram in the target. Moreover, sentences containing numbers are challenging for smaller models but are translated more accurately by T5-large.

Furthermore, we investigate the models' ability to adapt to the in-domain training vocabulary, specifically the pictogram terms encountered during training. By "training vocabulary," we refer to the unique words present in all sentences within the training set. We differentiate between the source vocabulary (words) and the target vocabulary (pictogram terms). For instance, the training set contains 20,177 sentences with 23,731 words in the source vocabulary and 4,354 pictogram terms in the target vocabulary. Similarly, the validation set includes 1,208 sentences with 3,558 words in the source vocabulary and 1,502 pictogram terms in the target vocabulary. Notably, the target vocabularies for the training and validation sets share 1,432 pictogram terms.

The T5 models are pre-trained on a vast amount of diverse text data; therefore, we expect the models to incorporate words seen during their pre-training in their predictions. During fine-tuning, the models should learn a new vocabulary of pictogram terms. We estimate the model's ability to do so by counting the words in the mutual vocabulary between the target sentences and the model predictions. We find that the T5-small and T5-base models include between 2,700 and 2,800 of the pictogram terms

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

encountered during training in their predictions. In comparison, the T5-large model appears to better adopt the target vocabulary seen during training, with approximately 3,500 mutual pictogram terms. This suggests that T5-large has a greater capacity to learn and utilize the target vocabulary effectively.

6.2. Speech-to-Picto

To solve the Speech-to-Picto task, we combine a pre-trained ASR model with the best model fine-tuned for Text-to-Picto translation. No training is involved in this process. Instead, we directly use the Whisper models for inference, hence, we do not make use of the training and validation sets for this task. As shown in Table 4, two different models from the Whisper family are used to produce transcripts from the audio of the test data. As expected, the larger Whisper model outperforms the smaller one, most likely due to higher-quality transcriptions.

Table 4
Model performance in terms of SacreBLEU, METEOR, and PictoER scores obtained for two different combinations of a Whisper ASR model (Speech-to-Text) and T5-large (Text-to-Picto). Only test set scores are included.

ASR model	MT model	SacreBLEU	METEOR	PictoER
whisper-small	T5-large	54.7	65.9	40.0
whisper-large		**62.9**	**73.4**	**29.5**

7. Conclusion and Future Work

In conclusion, the fine-tuned Google T5-large model exhibits strong performance in translating French text into appropriate sequences of pictograms. These promising results contribute to efforts to bridge the gap between AAC users and the broader society, facilitating effective communication. However, there is still room to improve the model's ability to generalize to unseen data and to reduce the generation of non-pictogram words.

Additionally, the Speech-to-Text-to-Picto solution, which utilizes Whisper to produce transcripts and the fine-tuned T5 model for translation, shows potential. Further refinement is needed to ensure accurate translations from spoken language to pictogram sequences.

In this study, the maximum number of tokens generated by the model was set to 64, since the longest sentences in the test set contained 62 words. Increasing this parameter could potentially improve predictions, depending on how tokens are generated with the T5 tokenizer. To enhance generalization on unseen data, techniques such as regularization or dropout could be employed, or the model could be trained on more diverse datasets. Furthermore, models fine-tuned for Text-to-Picto translation must adapt to a specialized vocabulary of pictogram terms. Future work could focus on further investigation and optimization of this in-domain adaptation.

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. Neither the European Union nor the Research Executive Agency can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to check grammar and spelling, paraphrase, and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] C. Syriopoulou-Delli, E. Gkiolnta, Effectiveness of different types of augmentative and alternative communication (aac) in improving communication skills and in enhancing the vocabulary of children with asd: a review, Review Journal of Autism and Developmental Disorders 9 (2021). doi:10.1007/s40489-021-00269-4.

[2] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of imageclef 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Grenoble, France, 2024.

[3] C. Macaire, D. Fabre, B. Lecouteux, D. Schwab, Overview of the 2025 imagecleftopicto task – investigating the generation of pictogram sequences from text and speech, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (nips 2017) 30 (2017).

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, Advances in Neural Information Processing Systems 33, Neurips 2020 33 (2020).

[7] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Naacl Hlt 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (2019) 4171–4186.

[8] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in Neural Information Processing Systems 33, Neurips 2020 33 (2020).

[9] C. Vaschalde, P. Trial, E. Esperança-Rodier, D. Schwab, B. Lecouteux, Automatic pictogram generation from speech to help the implementation of a mediated communication, 2018.

[10] V. Vandeghinste, L. Sevens, I. Schuurman, F. Van Eynde, Translating text into pictographs, Natural Language Engineering 23 (2017) 217–244. doi:10.1017/S135132491500039X.

[11] C. Macaire, C. Dion, D. Schwab, B. Lecouteux, E. Esperança-Rodier, Towards Speech-to-Pictograms Translation, in: Interspeech 2024, ISCA, Kos / Greece, Greece, 2024, pp. 857–861. URL: https://hal.science/hal-04687483. doi:10.21437/Interspeech.2024-490.

[12] 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: LREC-COLING 2024, Turin, Italy, 2024. URL: https://hal.science/hal-04534234.

[13] 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.

[14] M. Post, A call for clarity in reporting BLEU scores, in: O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, K. Verspoor (Eds.), Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 186–191. URL: https://aclanthology.org/W18-6319/. doi:10.18653/v1/W18-6319.

[15] B. Anand, T. J, S. Sai R, C. P, M. TT, SSN-MLRG at Text to Picto 2024: A BERT-Based Approach for Mapping French Sentences to Pictogram Terms, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/paper-135.pdf, notebook for the ImageCLEF Lab at CLEF 2024.

[16] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020. URL: http://dx.doi.org/10.18653/v1/2020.acl-main.645. doi:10.18653/v1/2020.acl-main.645.

[17] A. Elliah, A. Narayanan P, B. S, P. Mirunalini, Text-to-picto using lexical simplification, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/paper-146.pdf, notebook for the ImageCLEF Lab at CLEF 2024.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[19] M. Zampieri, P. Nakov, Y. Scherrer, Natural language processing for similar languages, varieties, and dialects: A survey, Natural Language Engineering 26 (2020) 595–612. doi:10.1017/S1351324920000492.

[20] A. Koushik, J. Morrison S, P. Mirunalini, J. A. R K, Text-to-picto using lexical simplification, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/paper-146.pdf, notebook for the ImageCLEF Lab at CLEF 2024.

[21] 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, 2023. URL: https://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[22] 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, Proceedings of the Twelfth Language Resources and Evaluation Conference (2020) 4218–4222. URL: https://aclanthology.org/2020.lrec-1.520/.

[23] C. Benzitoun, J.-M. Debaisieux, H.-J. Deulofeu, Le projet ORFÉO : un corpus d'étude pour le français contemporain, Corpus (2016). URL: http://journals.openedition.org/corpus/2936. doi:10.4000/corpus.2936, en ligne, mis en ligne le 15 janvier 2017, consulté le 24 mai 2025.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition

via large-scale weak supervision, 2022. URL: https://arxiv.org/abs/2212.04356. doi:10.48550/ARXIV.2212.04356.

[27] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), 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://aclanthology.org/W05-0909/.

[28] 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.

A. Loss Curves and T5-small and T5-base Model Performances

A.1. T5-small

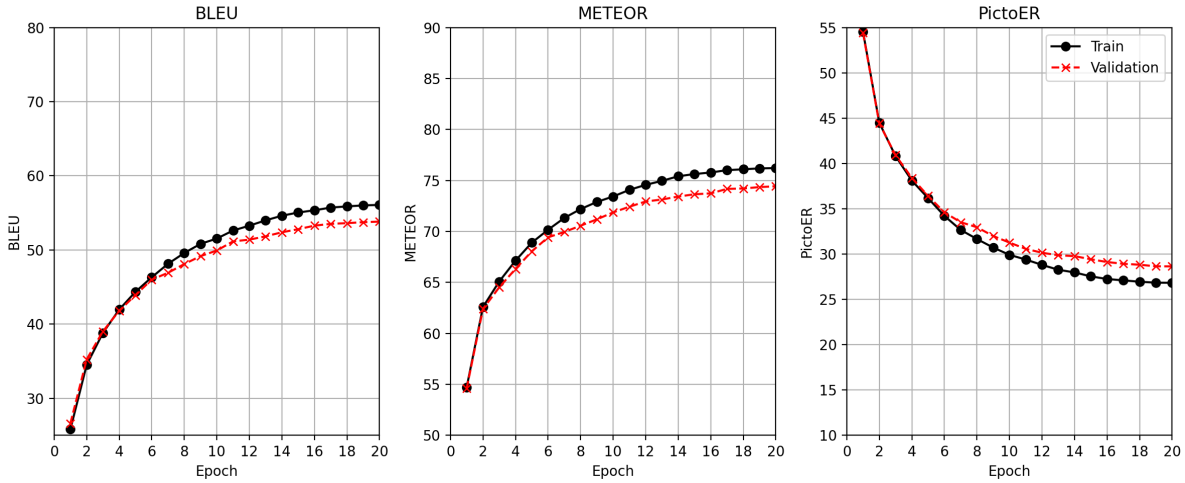


Figure 3: SacreBLEU, METEOR, and PictoER scores for ‘T5-small‘ finetuning for 20 epochs. The scales on the y-axes are identical to those in Fig. 4 for visual comparison.

A.2. T5-base

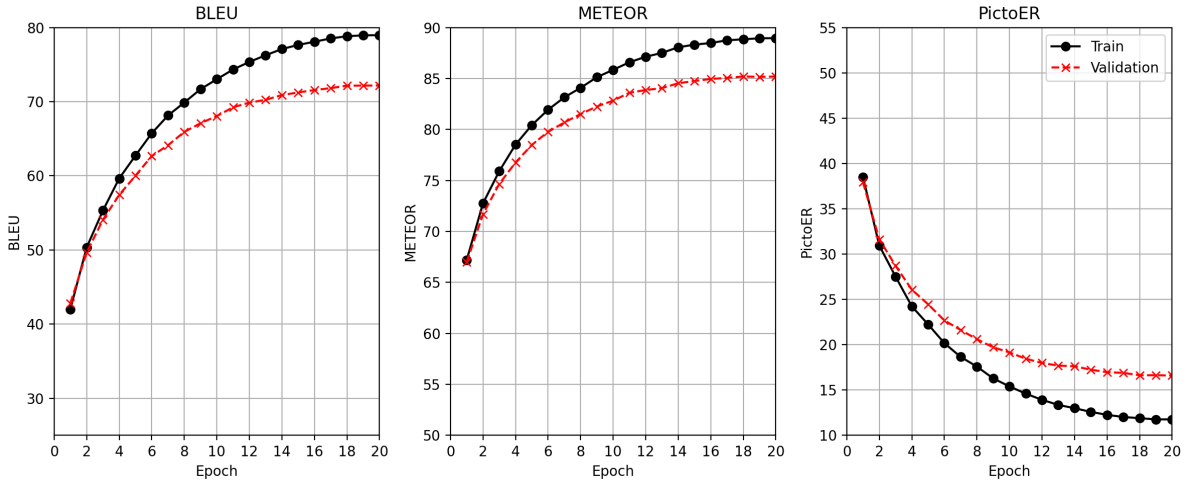


Figure 4: SacreBLEU, METEOR, and PictoER scores for ‘T5-base‘ finetuning for 20 epochs. The scales on the y-axes are identical to those in Fig. 3 for visual comparison.

A.3. Loss Curves

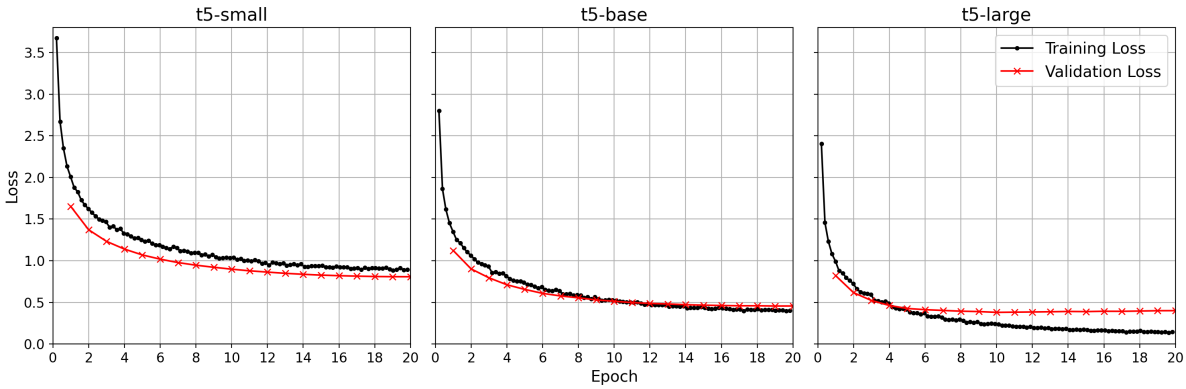


Figure 5: Training and validation loss for T5-small, -base, and -large while fine-tuning for 20 epochs.

B. Detailed Analysis of Generated Pictogram Sequences

This section presents the analysis of results obtained on the validation set with the different model sizes across four cases of linguistic analysis.

B.1. Case 1 - Generation of Non-pictograms Words

Case 1 (Table 5 and Figure 6) demonstrates the tendency of models to generate words that do not correspond to any existing pictogram. Although this tendency diminishes with increasing model size, the example in Table 5 shows that the word *"constater"* is still produced by the T5-large model, despite lacking an associated pictogram.

Table 5
Case 1 - Reference, target and sequences obtained with T5-small, -base and -large on the utterance of id **cefc-valibel-accFJ1r-16** (deviations from target are shown in bold, invalid pictograms are underlined).

src	non non ça non en france j'ai constaté à plusieurs reprises que on ne savait même pas dire si on était belge
tgt	non celle-là non au france passé me à plusieurs une_autre_fois prise_murale que nous même dire non si nous être belgique
T5-small	non celle-là non au france passé me **constater** à plusieurs **reprise** que nous dire non si nous être
T5-base	non celle-là non au france passé me **constater** à plusieurs **reprise** que nous **avoir** même dire non si nous être belgique
T5-large	non celle-là non au france passé me **constater** à plusieurs une_autre_fois prise_murale que nous **savoir non dire** si nous être belgique

B.2. Case 2 - Handling Past Tense

A limitation of the T5-small model was observed in its handling of past tense. As illustrated in Case 2 (Table 6 and Figure 7), although all generated pictograms are valid, the temporal aspect is lost in the output of T5-small. Both T5-base and T5-large correctly retain this temporal information.

Table 6
Case 2 - Reference, target and sequences obtained with T5-small, -base and -large on the utterance of id **common_voice_fr_18967443** (deviations from target are shown in bold).

src	son père y a ouvert un restaurant
tgt	passé son père ouvrir un restaurant
T5-small	son père **et avoir ouvert** un restaurant
T5-base	passé son père ouvrir un restaurant
T5-large	passé son père ouvrir un restaurant

B.3. Case 3 - Handling Names and Places

A specific feature of pictogram translation is that only some cities and countries have their own pictograms, otherwise a city will be translated by the generic pictogram "ville", a person by the generic pictogram "haut_du_corps", a pictogram that represents the upper body of a person. This rule is generally understood across all models. However, as illustrated in Case 3A (Table 7 and Figure 8), T5-small incorrectly interprets the city name "Saint-Paul" as a person, resulting in the pictogram "haut du corps". Both T5-base and T5-large provide the correct translation in this instance.

Case 3B (Table 8 and Figure 9) presents a more complex scenario involving the proper noun "Musikhochschule", the German word for "music school". The correct translation corresponds to the generic pictogram "association_à_but_non_lucratif" (non-profit organization). Both T5-small and

Table 7
Case 3A - Reference, target and sequences obtained with T5-small, -base and -large on the utterance of id **common_voice_fr_19658132** (deviations from target are shown in bold).

src	son président est joseph sinimalé maire de saint-paul
tgt	son président être haut_du_corps maire de ville
T5-small	son président être haut_du_corps maire de **haut_du_corps**
T5-base	son président être haut_du_corps maire de ville
T5-large	son président être haut_du_corps maire de ville

T5-base translate this term into the French "école_musicale", which, although semantically accurate, lacks a corresponding pictogram and is thus not a valid output. In contrast, T5-large successfully generates the appropriate pictogram. Nevertheless, the final two pictograms in T5-large's output are missing, indicating incomplete translation.

Table 8
Case 3B - Reference, target and sequences obtained with T5-small, -base and -large on the utterance of id **common_voice_fr_27080976** (deviations from target are shown in bold, invalid pictograms are underlined).

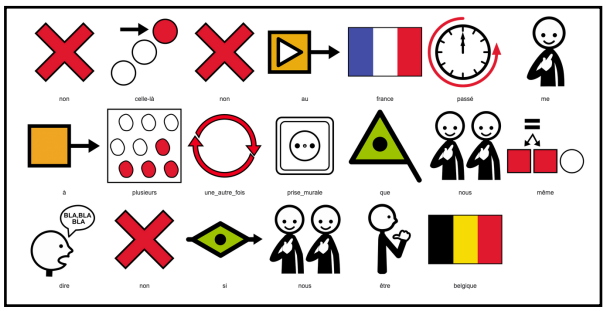
src	elle suit encore l'enseignement de johann sonnleitner à la musikhochschule de zürich
tgt	elle suivre une_autre_fois le formation_civique de haut_du_corps à le association_à_but_non_lucratif de ville
T5-small	elle suivre une_autre_fois le **enseignement** de haut_du_corps à le **école_musicale** de ville
T5-base	elle suivre une_autre_fois le **étudier** de haut_du_corps à le **école_musicale de ville**
T5-large	elle suivre une_autre_fois le formation_civique de haut_du_corps à le association_à_but_non_lucratif **[end is missing]**

B.4. Case 4 - Handling Numbers

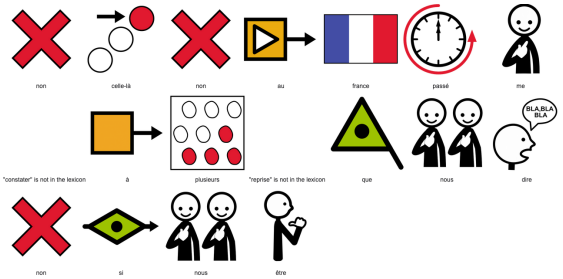
The final example, Case 4 (Table 9 and Figure 10), highlights a scenario in which even the T5-large model struggles to produce an accurate translation, particularly in handling numerical data and addresses. Although there is a noticeable improvement in translation quality with increasing model size in this example, one pictogram remains incorrectly translated in the T5-large output.

Table 9
Case 4 - Reference, target and sequences obtained with T5-small, -base and -large on the utterance of id **common_voice_fr_17801118** (deviations from target are shown in bold, invalid pictograms are underlined).

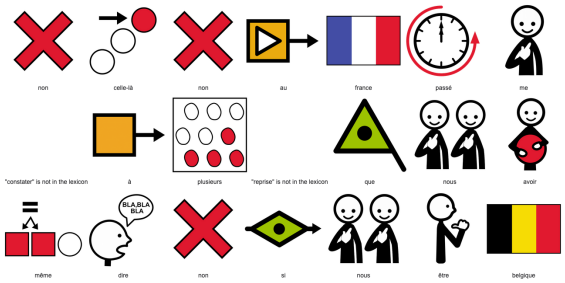
src	quatorze t square des tilleuls trente et un huit cent vingt pibrac
tgt	14 ville 30 et un 8 vingt ville
T5-small	**quatorze t carré de ville trente** et un **huit cent** vingt **pibrac**
T5-base	**quelqu'un toi carré de** trente et un 8 **20**
T5-large	14 **de** 30 et un 8 vingt ville



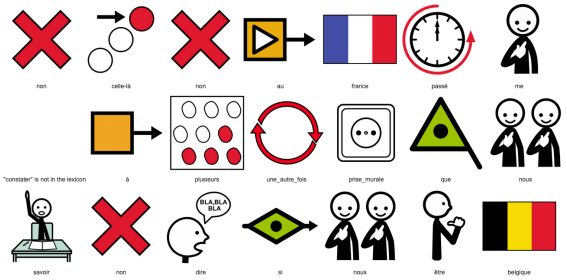
(a) tgt



(b) T5-small

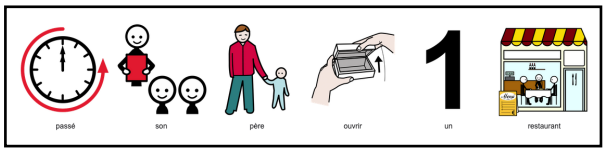


(c) T5-base



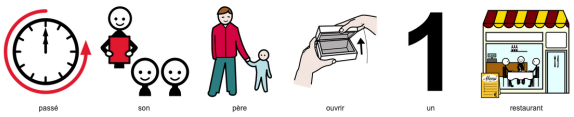
(d) T5-large

Figure 6: Case 1 - Pictogram sequences associated with target (framed in black) and sequences obtained with T5-small, -base and -large on the utterance of id **cefc-valibel-accFJ1r-16** (invalid pictograms are left blank).



(a) tgt

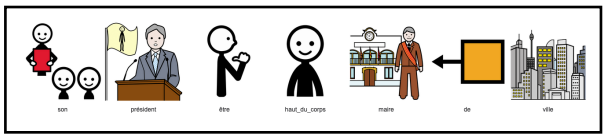
(b) T5-small



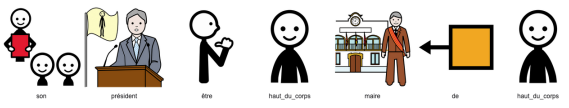
(c) T5-base

(d) T5-large

Figure 7: Case 2 - Pictogram sequences associated with target (framed in black) and sequences obtained with T5-small, -base and -large on the utterance of id **`common_voice_fr_18967443`**.



(a) tgt

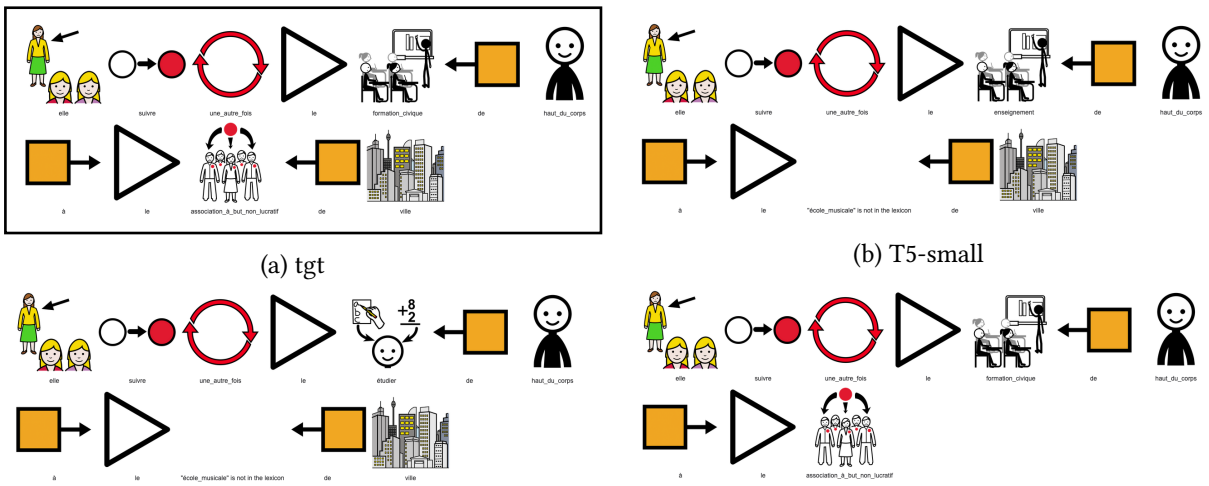


(b) T5-small

(c) T5-base

(d) T5-large

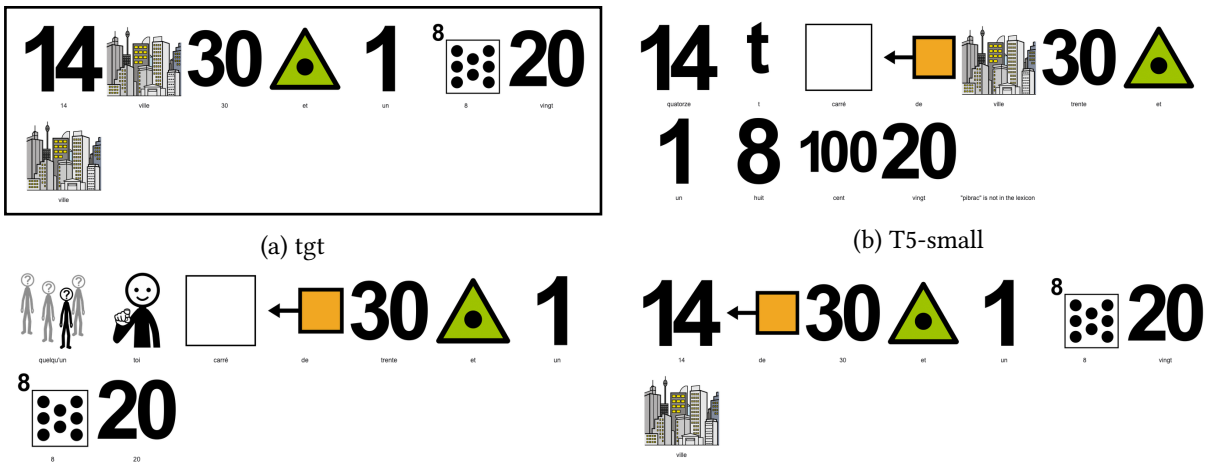
Figure 8: Case 3A - Pictogram sequences associated with target (framed in black) and sequences obtained with T5-small, -base and -large on the utterance of id **`common_voice_fr_19658132`**.



(a) tgt (b) T5-small

(c) T5-base (d) T5-large

Figure 9: Case 3B - Pictogram sequences associated with target (framed in black) and sequences obtained with T5-small, -base and -large on the utterance of id **`common_voice_fr_27080976`** (invalid pictograms are left blank).



(a) tgt (b) T5-small

(c) T5-base (d) T5-large

Figure 10: Case 4 - Pictogram sequences associated with target (framed in black) and sequences obtained with T5-small, -base and -large on the utterance of id **`common_voice_fr_17801118`** (invalid pictograms are left blank).