PICTOEDUCA: Building a Dataset for Spanish Text-to-Pictogram Generation

Alfonso Paredes Umeres and Marco Antonio Sobrevilla Cabezudo
Research Group on Artificial Intelligence, Pontificia Universidad Católica del Perú
{aparedesu, msobrevilla}@pucp.edu.pe

Abstract

We present PICTOEDUCA, the first large-scale Spanish text-to-pictogram dataset for augmentative and alternative communication (AAC), derived from primary educational materials and grounded in the ARASAAC pictogram repository. The dataset is released with a reproducible pipeline that combines automatic annotation with targeted expert correction, supporting scalable and high-quality corpus construction. We benchmark a rule-based system (ARAWORD) and neural models (T5, LLaMA) under direct text-to-pictogram and two-stage text-to-concept-to-pictogram settings. Results show that the rule-based system remains a strong baseline, while neural models benefit from explicit semantic abstraction, with the two-stage approach improving semantic coherence and reducing ambiguity. We further explore data selection strategies, demonstrating that combining domain similarity with a quality signal yields higher-quality silver data, reduces annotation effort, and improves model performance in low-resource regimes. PICTOEDUCA enables reproducible evaluation and advances Spanish text-to-pictogram research.

1 Introduction

Text-to-pictogram generation is a critical task in Augmentative and Alternative Communication (AAC), where natural language sentences are converted into sequences of pictograms to support comprehension and communication for individuals with cognitive or language impairments (Cabello et al., 2018; Schwab et al., 2020). Pictogram-based communication systems have been widely used in educational and clinical contexts, often leveraging tools such as ARAWORD in Spanish[1] or shared tasks such as ImageCLEFtoPicto in French[2], enabling systematic evaluation of translation and sequence prediction models (Norré et al., 2021; Koushik et al., 2024). While these efforts show the potential of data-driven and rule-based approaches, most resources target languages other than Spanish, use small corpora, or focus on AAC tools rather than structured datasets for benchmarking.

A key enabling resource for pictogram-based AAC systems is ARASAAC[3], an open and multilingual repository of pictograms that has been widely adopted in educational and clinical settings. ARASAAC pictograms serve as a shared visual vocabulary across languages and underpin a variety of tools and research efforts, including ARAWORD in Spanish and text-to-pictogram systems in French and Dutch (Schwab et al., 2020; Norré et al., 2021). The availability of a standardised pictogram inventory facilitates cross-lingual reuse, semantic alignment, and model comparability, while reducing the cost of symbol design and licensing. However, despite its broad adoption, ARASAAC does not provide large-scale, language-specific parallel corpora, leaving dataset construction and benchmarking as an open challenge, particularly for Spanish.

Despite Spanish being the second most spoken language worldwide, there is currently no large-scale Spanish dataset for text-to-pictogram generation. This absence limits the development of data-driven approaches, prevents reproducible evaluation, and constrains research in educational and accessibility applications for Spanish-speaking populations. In particular, creating high-quality, parallel sentence-pictogram corpora is challenging due to the labor-intensive nature of manual annotation and the subtle semantic nuances required for accurate pictogram representation (Bautista et al., 2017).

To address this gap, we present the first large-scale Spanish text-to-pictogram dataset, derived from Peruvian educational materials. The dataset is complemented by a reproducible construction

[1] Available at https://aulaabierta.arasaac.org/araword_inicio.
[2] Available at https://www.imageclef.org/2025.
[3] Available at https://arasaac.org/.

Findings of the Association for Computational Linguistics: ACL 2026, pages 34816–34828

pipeline that integrates automatic annotation with targeted expert correction and quality control, providing a generalisable framework for creating comparable resources in other languages or domains.

We benchmark a range of text-to-pictogram generation strategies and models, and investigate a data selection strategy to prioritise high-quality, informative instances. This approach enables the construction of compact, high-quality 'silver' seed subsets, facilitates efficient annotation, and improves model performance. Our results establish baselines and shed light on the impact of selective annotation when scaling text-to-pictogram datasets. In summary, we make the following contributions:

- We release PICTOEDUCA, the first large-scale Spanish text-to-pictogram corpus derived from educational materials, supporting reproducible evaluation.
- We provide a systematic evaluation of rule-based and neural generation strategies.
- We propose a pipeline for selecting, simplifying and annotating sentences that can be easily adapted to other domains and languages.
- We demonstrate how selective annotation can accelerate corpus creation, identify a high-quality seed dataset, and potentially improve model performance.

2 Related Work

Spanish Work In Spanish, ARAWORD[4] is a widely used AAC tool displaying ARASAAC pictograms alongside text, while AraTraductor uses syntactic and morphological Natural Language Processing techniques to map sentences to pictograms (Bautista et al., 2017). Despite these resources, large-scale, parallel Spanish text-to-pictogram corpora for data-driven models remain unavailable, a gap addressed by our dataset.

Other Languages In French, the ImageCLEFtoPicto shared task[5] provides text- and speech-to-pictogram pairs, evaluated with metrics such as BLEU, METEOR, and PictoER (Macaire et al., 2025). Dutch systems and Arasaac-WN further illustrate cross-lingual and semantic alignment approaches linking text to pictograms (Vandeghinste et al., 2017; Sevens et al., 2015; Schwab et al., 2020; Norré et al., 2021).

Multilingual Prediction Models Transformer-based models such as PictoBERT and BERTimbau (Souza et al., 2020) variants demonstrate the value of sequence prediction for pictogram generation in multiple languages (Pereira et al., 2022, 2024). Other work in Bengali and multimodal French corpora highlight the importance of aligned datasets for bidirectional text-to-AAC generation and speech-to-pictogram tasks (Karmakar and Sinha, 2024; Macaire et al., 2024).

3 PICTOEDUCA

3.1 Data Sources

The corpus was derived from official Peruvian primary school textbooks published by the Ministry of Education of Peru (MINEDU) and distributed through the Perú Educa platform and the Institutional Repository[6]. A total of 18 textbooks were collected from the subjects Communication and Science and Technology, targeting pupils from third to sixth grade of primary education.

The selected materials span the years 2020-2023 and include both regular and rural educational modalities. All resources are publicly available and were chosen to ensure linguistic consistency, educational relevance, and age-appropriate content. The diversity of years and modalities allows the corpus to capture a broad range of instructional styles and vocabulary commonly encountered in Peruvian primary education.

3.2 PICTOEDUCA Building Pipeline

Figure 1 illustrates the procedure used to construct the PICTOEDUCA dataset.

3.2.1 Preprocesing and Filtering

Sentences were automatically extracted from PDF documents using custom web scraping. Following extraction, all sentences were manually validated to ensure grammatical correctness, semantic coherence, and relevance to educational contexts. This initial step yielded 27,650 sentences.

A subsequent cleaning phase removed duplicated entries, sentences containing non-Spanish tokens, and fragments lacking sufficient contextual meaning (e.g. cut segments), reducing the corpus to 21,962 unique sentences.

[4] Available at https://aulaabierta.arasaac.org/araword_inicio.

[5] Available at https://www.imageclef.org/2025.

[6] Available at https://www.perueduca.pe/#/home/materiales-educativos.

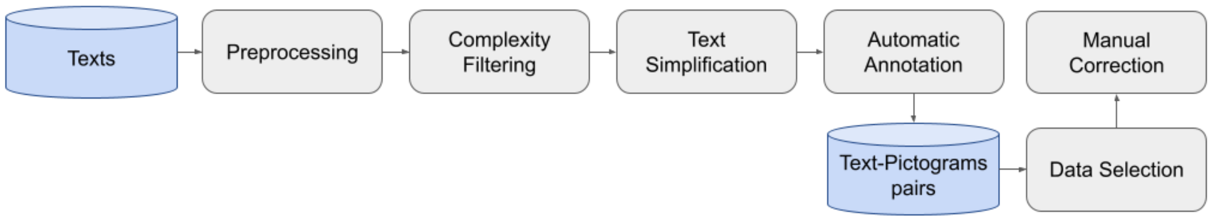

Figure 1: Pipeline for constructing the PICTOEDUCA dataset

Given the target audience of children and individuals with communication difficulties, sentences were categorised according to their linguistic complexity. We adopted the proposal of Vásquez-Rodríguez et al. (2022) for readability assessment in Spanish. Using this framework, sentences were classified into simple and complex categories.

The resulting distribution comprised 17,626 simple sentences and 4,336 complex sentences. Only simple sentences were retained for further processing, as they typically exhibit shorter length, lower lexical density, and more predictable syntactic structures, all of which are beneficial for pictographic translation tasks.

3.2.2 Automated Text Simplification

To further enhance accessibility, the selected "simple" sentences were automatically simplified using GPT-4, which was prompted to enforce easy-to-read principles, a maximum of ten words, and vocabulary suitable for 7–8-year-old children (see Figure 4, Appendix A). This step reduced residual syntactic and lexical complexity while preserving meaning. After cleaning and deduplication, the corpus was reduced to 16,319 simplified sentences.

3.2.3 Annotation Process

Each sentence was annotated with a sequence of pictogram identifiers from the ARASAAC system[7]. ARASAAC provides an open-access pictogram repository, exposed via public APIs[8], containing metadata for over 13,500 pictograms, including identifiers and associated keywords.

Automatic annotation was performed using the static rule-based system ARAWORD[9]. For each sentence, tokens were lemmatised and matched against the ARASAAC dictionary. When a direct match was unavailable, fallback strategies were applied, such as using the lemma form or preserving the original word as metadata. Compound pictograms were detected and substituted when applicable. The final output for each instance consisted of the original sentence, the pictogram identifier sequence, and associated metadata. The algorithm is described in Appendix B.

In addition to the automatically annotated corpus, two manually labelled subsets were created for validation and test. First, 1,111 sentences were sampled from the corpus and annotated by expert psychologists specialised in child language, behaviour, and augmentative and alternative communication (AAC). To support this process, a dedicated web-based annotation tool was developed, allowing experts to select pictograms directly from the ARASAAC repository.

Second, 835 sentences were manually extracted from ARASAAC educational materials and annotated following the same linguistic and pedagogical criteria. These two subsets were combined and split evenly into validation and test sets, each containing 973 sentences. All manually annotated instances were excluded from the automatically annotated data, leaving 15,208 sentences for training.

4 Experimental Setup

4.1 Task Definition

The task consists of translating a Spanish sentence into a sequence of pictogram identifiers from the ARASAAC repository. Given an input sentence $x = (w_1, \ldots, w_n)$, the goal is to generate an ordered sequence $y = (pict_1, \ldots, pict_n)$, where each $pict_n$ corresponds to a pictogram identifier that visually represents part of the sentence meaning. Figure 2 shows an example.

The task is framed as a sequence generation problem and evaluated at the sentence level. Two task formulations are considered:

- Direct text-to-pictogram translation: where the model predicts pictogram identifiers directly from text.

[7]Available at https://arasaac.org/.
[8]Available at https://arasaac.org/developers/api.
[9]Available at https://aulaabierta.arasaac.org/araword_inicio

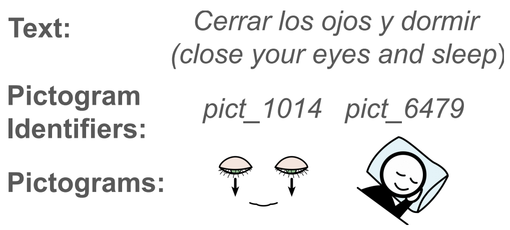

Figure 2: Input text, pictogram identifiers and pictograms (images) for the sentence "Cerrar los ojos y dormir".

- Two-stage text-to-concept-to-pictogram translation: where an intermediate sequence of textual concepts is first generated and then mapped to pictograms.

4.2 Modeling Approaches

To address the task, three different modelling strategies were implemented: a rule-based baseline, a large language model using prompting, and a neural sequence-to-sequence model.

4.2.1 Rule-Based Baseline: ARAWORD

As a baseline, we employ ARAWORD, a rule-based system designed to translate Spanish sentences into ARASAAC pictograms. ARAWORD relies on lemmatisation, dictionary matching, and handcrafted linguistic rules to map words or lemmas to pictogram identifiers.

4.2.2 Large Language Model: LLaMA-8B

The second approach uses LLaMA 3.1–8B-Instruct (Grattafiori et al., 2024) under a two-stage text-to-concept-to-pictogram strategy: the model first predicts key concepts from each sentence, which are then mapped to pictogram identifiers.

The prompt (Figure 5, Appendix C) guides the model to extract visually representable actions, entities, attributes, and relations, following AAC principles (e.g., implicit subjects, generic nouns). The resulting concepts are mapped to ARASAAC pictograms via keyword matching.

To resolve multiple candidate pictograms, a multilingual CLIP model (Reimers and Gurevych, 2019)[10] computes text–image similarity, selecting the pictogram that best aligns with each concept.

4.2.3 Sequence-to-Sequence Model: T5

The third approach employs a neural encoder–decoder architecture based on the Spanish T5 Base model (Raffel et al., 2020; Araujo et al., 2024), fine-tuned on the annotated corpus under both the direct text-to-pictogram and two-stage text-to-concept-to-pictogram settings.

In the direct setting, the model generates pictogram identifiers as target tokens, while in the concept-based configuration it predicts textual concepts, which are mapped to pictograms via the same CLIP-based selection used for LLaMA.

In addition, for direct setting, we extend model's vocabulary to include all ARASAAC identifiers. Training used a batch size of 16, a learning rate of 1×10^{-5}, 40 epochs, and a weight decay of 0.01.

4.3 Automatic Evaluation

System outputs were evaluated against expert-annotated references using standard sequence generation metrics:

- BLEU (Papineni et al., 2002), to measure n-gram overlap between predicted and reference pictogram sequences.
- chrF++ (Popović, 2017), which captures character-level similarity and is more tolerant of near-miss pictogram identifiers.
- n-gram precision (1–4), to analyse performance degradation as sequence length increases.

Table 1 presents the automatic evaluation results for the proposed text-to-pictogram approaches. The rule-based system ARAWORD achieves the best overall performance, obtaining the highest BLEU (0.3741), chrF++ (51.46), and n-gram precision scores. These results confirm ARAWORD as a strong baseline, particularly in scenarios where a well-defined pictographic vocabulary and deterministic mappings are available, highlighting the relevance of rule-based methods in AAC settings.

Neural models benefit from a two-stage pipeline that separates semantic interpretation (Text-to-Concept) from pictogram selection (Concept-to-Pictogram). This is evident for T5, whose BLEU score increases from 0.2977 in the direct setting to 0.3218 in the two-stage configuration, alongside a substantial improvement in chrF++ (42.95 to 48.23). This decomposition appears well suited to pretrained text-based models, as the direct approach requires learning and sequencing pictogram identifiers, which likely demands more data. Aligning pictogram embeddings with pretrained textual representations may help alleviate this issue.

[10]Available at https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

Within the two-stage setting, LLaMA-8B achieves its highest BLEU score (0.3452) under 1-shot prompting. While additional in-context examples do not consistently improve BLEU, they yield moderate gains in chrF++, suggesting improved selection of semantically related pictograms even when exact label matches are not achieved (e.g., alternative pictograms for "*yo*" - Figure 3). However, despite strong unigram precision, LLaMA-8B shows a marked drop in higher-order n-gram precision, indicating difficulties in generating coherent longer pictogram sequences. Future work may explore larger or specialised models and structure-aware decoding strategies to address this limitation.

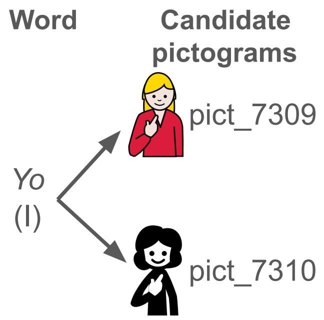

Figure 3: Candidate pictograms and its corresponding pictogram ids for the word "Yo" (I).

4.4 Human Evaluation

To assess semantic adequacy and interpretability, we conducted a human evaluation with eight domain experts, including psychologists, speech therapists, and special education professionals experienced in AAC. A random sample of 50 test sentences was selected, and for each instance, evaluators rated the correspondence between the input sentence and the generated pictogram sequence on a five-point Likert scale (1 = very poor, 5 = very good). The evaluation explicitly targets semantic adequacy rather than strict sequence matching.

Evaluation Interface. The evaluation was performed through a custom annotation interface designed to facilitate consistent judgement across annotators. Each evaluation instance presented the original sentence alongside the generated pictogram sequence rendered visually using ARASAAC symbols. Annotators were asked to assess the overall correspondence between the sentence and the pictogram sequence, considering semantic adequacy. No gold reference was shown, encouraging evaluators to rely on their professional judgement of communicative adequacy. The interface enforced independent scoring and randomized instance order to mitigate ordering effects. Additional details and screenshots of the interface are provided in Appendix 6.

Inter-annotator agreement. We measure inter-annotator agreement using Krippendorff's alpha with an ordinal distance function, obtaining $\alpha = 0.2172$, which indicates low agreement. Given the subjective and multimodal nature of the task, this result reflects substantial variability in how semantic adequacy and acceptable pictogram realisations are interpreted.

To better understand this variability, we performed annotator- and item-level analyses. At the annotator level, we computed leave-one-out agreement (LOO-α), deviation-from-consensus, and pairwise correlations (showing mean and standard deviation)[11]. The analysis revealed two annotators with systematically higher disagreement relative to the rest of the group. After excluding them, agreement increased to $\alpha = 0.3798$. While still modest, this improvement indicates that part of the disagreement is attributable to annotator inconsistency, though substantial variability remains.

At the item level, we analysed dispersion across annotations using multiple measures (standard deviation, range, entropy, and deviation from the median). Disagreement is not uniformly distributed, but instead concentrated in a subset of sentences exhibiting higher semantic ambiguity, multiple plausible pictogram realisations, or abstract content.

To further disentangle annotator effects from intrinsic task difficulty, we conducted a stability analysis of high-disagreement items. Specifically, we identified the top-30 items with highest dispersion before and after removing the two most divergent annotators. Notably, 14 out of the original 30 items remained in the top-30 set, indicating substantial overlap. This partial stability suggests that a significant portion of disagreement is intrinsic to the task, reflecting genuine subjectivity and ambiguity rather than solely annotator noise. Two examples of the dataset with high disagreement are shown in Figure 7 at Appendix F.

Human evaluation results. Table 2 reports the human evaluation results after removing the two most inconsistent annotators, which we consider

[11]The results of this analysis can be found in Table 6 at Appendix E

Setting	Approach	BLEU	Precision-1	Precision-2	Precision-3	Precision-4	chrF++
TEXT-TO-PICTOGRAM	ARAWORD	**0.3741**	0.6075	0.4721	**0.3096**	**0.2205**	**51.46**
	T5	0.2977	0.5699	0.4107	0.2187	0.1534	42.95
TEXT-TO-CONCEPT-TO-PICTOGRAM	T5	0.3218	0.5766	0.4273	0.2494	*0.1745*	*48.23*
	LLaMA-8B 0-SHOT	0.3061	0.6447	0.4563	0.2204	0.1354	41.43
	LLaMA-8B 1-SHOT	*0.3452*	**0.6562**	**0.4790**	*0.2581*	*0.1751*	44.60
	LLaMA-8B 2-SHOT	0.3195	0.6313	0.4526	0.2330	0.1565	44.53
	LLaMA-8B 3-SHOT	0.3229	0.6278	0.4522	0.2373	0.1614	45.16

Table 1: Automatic evaluation results. **Bold** values indicate the overall highest scores, while *italic* values indicate the highest scores within the Text-to-Concept-to-Pictogram setting. The results reported for T5 and LLaMA are based on only 1 execution.

the most reliable estimate of system performance (results before annotator filtering can be found in Table 7 at Appendix G).

In general, ARAWORD achieves the highest score (3.55), confirming its strength as a rule-based baseline. Among neural approaches, T5 benefits substantially from the Text-to-Concept-to-Pictogram setting (3.21), clearly outperforming the direct Text-to-Pictogram variant (2.47). This result indicates that introducing an explicit conceptual layer helps reduce semantic ambiguity in pictogram generation. LLaMA obtains a moderate score (2.74), suggesting reasonable performance but still below concept-based approaches on T5.

Approach	Avg. Score
ARAWORD	**3.55***
T5 (Text-to-Pictogram)	2.47
T5 (Text-to-Concept-to-Pictogram)	3.21
LLaMA-8B-1-SHOT	2.74

Table 2: Human evaluation scores after removing two annotators. *Indicates statistically significant difference.

Statistical analysis confirms that these differences are significant. A Friedman test reveals overall variation across systems ($\chi^2 = 21.62$, $p < 0.05$), and post-hoc Wilcoxon tests ($p < 0.05$) with Holm correction show that ARAWORD significantly outperforms all other approaches.

Finally, we observe a moderate system-level correlation between automatic metrics and human judgements. While metrics such as BLEU and chrF++ capture broad performance trends, they fail to reflect finer-grained communicative adequacy. In particular, models with strong lexical overlap (e.g., LLaMA) receive lower human ratings, suggesting that n-gram metrics overestimate performance by prioritizing local overlap over global semantic coherence. Among the evaluated metrics, chrF++ shows the closest alignment with human judgement, yet none adequately capture acceptable semantic variation in AAC, highlighting the need for more semantically grounded evaluation methods.

5 Impact of Data Selection on Annotation and Model Performance

Annotating sentences with corresponding pictogram sequences is labour-intensive, and although automatic systems such as ARAWORD enable scalable annotation, they often introduce noise and fail to capture fine-grained semantic distinctions. Our initial corpus of 15,208 automatically annotated sentences therefore contained inconsistencies that made exhaustive manual correction impractical. In this section, we investigate whether data selection can mitigate these limitations by guiding annotation effort towards the most informative instances. Specifically, we examine *whether data selection strategies are effective for identifying high-quality silver data*, *accelerating the annotation process* by prioritising impactful examples, and ultimately *contributing to measurable improvements in downstream model performance*. To this end, we adopt a pool-based active learning framework to select candidate instances for expert review, with the goal of improving annotation quality while reducing manual effort and constructing a more reliable text-to-pictogram dataset.

5.1 Selection Criteria

We defined two complementary criteria to estimate the informativeness of a sentence for correction:

Domain Similarity via Feature Decay Algorithms (FDA) : To ensure that reviewed instances were representative of the intended application domain (educational texts for children), we applied the Feature Decay Algorithm (FDA) (Biçici and

Yuret, 2015). FDA ranks candidate sentences based on their n-gram overlap with a reference set (S_{seed}) drawn from the validation corpus. The score for each candidate sentence s is computed as:

$$score(s, S_{seed}, L) = \frac{\sum_{ngr \in \{s \cap S_{seed}\}} 0.5^{C_L(ngr)}}{length(s)}$$

Where ngr denotes an n-gram, and $C_L(ngr)$ is the count of how often the n-gram has already appeared in the selected set L. This scoring prioritizes sentences that introduce informative, diverse content relative to the validation domain.

Translation Quality Estimation via X-CLIP (TQE) : As gold-standard pictogram references are unavailable for most of the training data, we propose a proxy metric to estimate translation quality. Pictogram sequences are modelled as visual streams and compared to sentence meaning using X-CLIP (Ni et al., 2022), a multimodal transformer trained for video–text alignment. As X-CLIP primarily operates in English, Spanish sentences are first translated into English using GPT-4. Cosine similarity between the translated text and the pictogram embeddings provides a weak semantic alignment score, with higher values indicating more faithful pictographic translations.

To combine both metrics into a unified selection mechanism, we defined a composite informativeness score for each instance:

$$score(s) = FDA(s).TQE(s)$$

This multiplicative formulation ensures that selected examples are both domain-relevant and pictographically meaningful, filtering out low-impact or low-quality candidates.

5.1.1 Data Selection for better silver data

We evaluate data selection strategies using a Spanish T5 model trained for direct text-to-pictogram generation. The initial training corpus consists of 15,208 automatically annotated sentences. From this pool, we construct subsets of varying sizes (1,000, 3,000, 5,000, and 7,000 sentences) using three sampling strategies:

- **Random**: sentences sampled uniformly at random from the base corpus (baseline);
- **Similarity**: sentences selected based on domain similarity using FDA;
- **Similarity + Quality**: sentences selected using a combination of FDA-based similarity and TQE-based quality estimation.

Table 3 shows the results of the experiments. Across all settings, performance improves with increasing data size, reflecting the benefits of additional training signal. For example, BLEU scores rise from approximately 0.009–0.019 with 1,000 examples to above 0.23 with 7,000 examples, while chrF++ increases from around 20 to over 36.

The selection strategy has a pronounced effect in low-data regimes. Random sampling consistently yields the weakest results, while similarity-based selection provides moderate gains by prioritising domain-relevant examples. The strongest improvements are observed when similarity is combined with quality-based filtering. With only 1,000 examples, Precision-1 increases from 0.332 under Random sampling to 0.481 with Similarity + Quality, demonstrating the value of careful data curation.

As the subset size grows, the performance gap between Similarity and Similarity + Quality narrows, indicating diminishing returns from quality filtering in higher-resource settings. At 7,000 examples, BLEU scores are nearly identical for both strategies (0.228 vs. 0.230), although both outperform Random selection. Overall, these results suggest that quality-aware data selection is particularly effective for constructing high-quality initial training sets, while larger datasets naturally reduce the marginal benefit of additional filtering.

5.1.2 Annotation Efficiency

We estimate annotator effort by computing the word-level edit distance between the model-generated pictogram sequences and their human-corrected versions. This metric reflects the number of insertions, deletions, and substitutions required for a sentence to be usable in practice, with higher values indicating greater annotation effort.

To analyse annotation efficiency under realistic selection conditions, we consider a subset of 1,000 instances selected using two strategies: *Domain Similarity* and *Domain Similarity + Quality*. Since data selection methods typically prioritise examples ranked as most informative, we focus on the *last 250 instances* selected by each strategy. These instances are expected to be the least similar to the validation set; however, under the Similarity + Quality strategy, they are also constrained to satisfy a minimum quality criterion. This setup allows us

Subset Size	Strategy	BLEU	Precision-1	Precision-2	Precision-3	Precision-4	chrF++
1000	Random	0.0092	0.3322	0.1768	0.0004	0.0003	20.46
	Similarity	0.0117	0.3034	0.1615	0.0009	0.0004	19.07
	Similarity + Quality	0.0195	0.4813	0.2695	0.0018	0.0006	20.63
3000	Random	0.0502	0.5130	0.2866	0.0082	0.0053	25.80
	Similarity	0.0523	0.5064	0.2823	0.0091	0.0058	26.84
	Similarity + Quality	0.0594	0.5079	0.2842	0.0114	0.0075	27.28
5000	Random	0.1418	0.5189	0.3123	0.0601	0.0415	29.87
	Similarity	0.1617	0.5184	0.3186	0.0761	0.0543	31.01
	Similarity + Quality	0.1631	0.5178	0.3189	0.0776	0.0552	31.21
7000	Random	0.2069	0.5352	0.3472	0.1182	0.0835	34.31
	Similarity	0.2283	0.5421	0.3607	0.1403	0.0989	36.35
	Similarity + Quality	0.2303	0.5416	0.3614	0.1427	0.1008	36.58

Table 3: Automatic evaluation across data selection strategies and subset sizes

to assess which strategy better identifies examples suitable for expert review.

Table 4 reports the distribution of the edit distances for both strategies. Results show that incorporating quality information consistently reduces annotator effort. The proportion of examples requiring no changes increases to 20% under Similarity + Quality, compared to 14% with Similarity alone, while high-effort cases (edit distance > 0.75) decrease from 12.8% to 9.2%. Mid-range edit distances (0.25–0.75) remain comparable across strategies, indicating that most annotation effort is concentrated on moderately challenging examples.

Overall, these findings suggest that combining domain similarity with a quality-aware signal more effectively filters out low-quality instances, reducing the number of examples that require more manual correction and improving annotation efficiency.

Edit Distance Range	Similarity (%)	Similarity + Quality (%)
0.0 (no changes)	14.0	**20.0**
(0.0, 0.25]	10.8	14.0
(0.25, 0.5]	38.4	37.6
(0.5, 0.75]	24.0	19.2
>0.75	12.8	**9.2**

Table 4: Distribution of edit distances under different data selection strategies.

5.1.3 Impact of Manual Correction on Model Performance

To assess how expert intervention affects downstream model performance, we compare models trained on automatically annotated data with those trained on manually corrected instances selected under different data selection strategies.

Using the same subset analysed for annotation efficiency, Table 5 compares the Similarity and Similarity + Quality strategies under both automatic ("auto") and human-corrected ("corrected") conditions. Model performance is reported using Precision-1, Precision-2, and chrF++[12].

For the Similarity-based strategy, manual correction leads to consistent improvements across all metrics, with Precision-1 increasing from 0.3085 to 0.3793, Precision-2 from 0.1639 to 0.2051, and chrF++ from 19.03 to 19.70. This indicates that similarity-driven selection surfaces a higher proportion of structurally noisy or misaligned annotations, for which human intervention directly improves surface-level alignment and lexical consistency.

In contrast, for the Similarity + Quality strategy, manual correction results in a slight decrease across metrics. Rather than indicating a negative effect of human intervention, this behaviour suggests diminishing returns when a strong quality signal is already present during selection. In this setting, remaining errors tend to be more subtle and semantic in nature, and manual corrections often prioritise communicative adequacy over exact surface overlap. Given the partial nature of correction, this may also introduce minor inconsistencies within an otherwise highly regularised subset, which are not well captured by single-reference, n-gram–based metrics such as Precision and chrF++.

Overall, these results highlight that manual correction is most beneficial when applied to data selected by weaker heuristics, where annotation noise is more pronounced. When quality-aware selection already constrains the error space, the marginal gains of partial human correction diminish and may not be reflected by automatic evaluation metrics.

[12]BLEU and Precision-3 and Precision-4 are omitted due to near-zero values and limited interpretability.

Strategy	Precision-1	Precision-2	chrF++
Similarity (auto)	0.3085	0.1639	19.03
Similarity (corrected)	**0.3793**	**0.2051**	**19.70**
Similarity + Quality (auto)	**0.4742**	**0.2641**	**20.34**
Similarity + Quality (corrected)	0.4091	0.2222	**20.31**

Table 5: Automatic Evaluation for Simility and Similarity + Quality selection strategies before and after correct 250/1000 instances.

6 Conclusion and Future Work

We introduced PICTOEDUCA, a large-scale Spanish dataset for text-to-pictogram generation grounded in ARASAAC, along with a reproducible pipeline that combines automatic annotation with targeted expert correction to support scalable dataset construction for AAC.

We benchmarked rule-based and neural approaches under direct and two-stage text-to-concept-to-pictogram settings. Automatic and human evaluations show that the rule-based system remains a strong baseline, while neural models consistently benefit from explicit semantic abstraction, with the two-stage formulation reducing ambiguity and improving performance.

Finally, we showed that quality-aware data selection improves both annotation efficiency and model performance, particularly in low-resource settings. Combining domain similarity with a proxy quality signal yields higher-quality silver data, reduces annotation effort, and enables stronger models with fewer training examples, with diminishing gains as data size increases.

Future work includes expanding the dataset to additional domains, linguistic variations, and expert annotations to improve generalisation. We also plan to develop evaluation metrics tailored to pictogram sequences that better capture semantic relations and multiple valid realisations, extend the approach to other languages using multilingual resources, and further fine-tune models with human feedback to improve robustness. The dataset provides a reproducible platform for advancing both model development and evaluation in automatic pictogram translation, supporting scalable and socially impactful applications.

Limitations

Despite the contributions of the benchmark and the evaluated models, several limitations remain:

Dataset characteristics : PICTOEDUCA is restricted to Spanish texts from Peru and focuses primarily on educational content, limiting coverage of other domains. In addition, each PICTOEDUCA's instance has a single reference, which may underestimate model performance and limit evaluation variability.

Evaluation metrics : Standard metrics such as BLEU and chrF++ only partially reflect human judgement, highlighting the need for more semantically grounded evaluation methods.

Neural model constraints : Models struggle to learn the "pictogram language" when it contains unseen identifiers, highlighting the need for strategies to induce embeddings for these tokens. One potential approach is to initialise pictogram embeddings using the embeddings of their definitions or keywords provided by ARASAAC.

Data selection experiments : Experiments were conducted on small subsets (250 from 1,000 instances), limiting conclusions on the effect of selective annotation to better assess model performance improvements (from subsection 5.1.3). Furthermore, data selection strategies leveraging domain representativeness (FDA) and quality-based filtering (TQE) proved effective. However, these approaches rely on proxy measures—particularly X-CLIP, which is designed for video–text alignment and may not perfectly capture the semantic fidelity of pictogram sequences.

Human evaluation scope. The human evaluation was conducted on a subset of 50 sentences assessed by eight experts, which may limit the generalisability of the findings. Future work should consider larger-scale evaluations to provide a more robust estimate of model performance. Additionally, more fine-grained evaluation protocols are needed to better understand which aspects of generation are captured beyond overall semantic adequacy. Finally, including children with special communication needs could provide stronger evidence of real-world effectiveness, although the current evaluation focused primarily on adequacy rather than communicative outcomes.

Information About Use of Artificial Intelligence (AI) Assistants

AI-based tools were used solely for language polishing and stylistic refinement of the paper. They were not used to generate research ideas, experimental results, analyses, datasets, or conclusions.

All content, experimental design, and interpretations were produced and verified by the authors, who take full responsibility for the paper.

References

Vladimir Araujo, Maria Mihaela Trusca, Rodrigo Tufiño, and Marie-Francine Moens. 2024. Sequence-to-sequence Spanish pre-trained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14729–14743, Torino, Italia. ELRA and ICCL.

Susana Bautista, Raquel Hervás, Agustín Hernández-Gil, Carlos Martínez-Díaz, Sergio Pascua, and Pablo Gervás. 2017. Aratraductor: text to pictogram translation using natural language processing techniques. In *Proceedings of the XVIII International Conference on Human Computer Interaction*, Interacción '17, New York, NY, USA. Association for Computing Machinery.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):339–350.

L. Cabello, E. Lleida, J. Simon, A. Miguel, and A. Ortega. 2018. Text-to-Pictogram Summarization for Augmentative and Alternative Communication. *Procesamiento de Lenguaje Natural*, 61:15–22.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Piyali Karmakar and Manjira Sinha. 2024. Aiding non-verbal communication: A bidirectional language agnostic framework for automating text to AAC generation. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 324–331, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Avaneesh Koushik, Jithu Morrison, P Mirunalini, and 1 others. 2024. A transformer based approach for text-to-picto generation. In *Notebook for the ImageCLEF Lab at CLEF 2024*, pages 1656–1661.

Cécile Macaire, Chloé Dion, Jordan Arrigo, Claire Lemaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, and Didier Schwab. 2024. A multimodal French corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 839–849, Torino, Italia. ELRA and ICCL.

Cécile Macaire, Diandra Fabre, Benjamin Lecouteux, and Didier Schwab. 2025. Overview of the 2025 ImageCLEFtoPicto Task -Investigating the Generation of Pictogram Sequences from Text and Speech Notebook for the ImageCLEF Lab at CLEF 2025. In *Springer Lecture Notes in Computer Science (LNCS)*, Madrid, Spain.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models fornbsp;general video recognition. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, page 1–18, Berlin, Heidelberg. Springer-Verlag.

Magali Norré, Vincent Vandeghinste, Pierrette Bouillon, and Thomas François. 2021. 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)*, pages 1050–1059, Held Online. INCOMA Ltd.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jayr Pereira, Rodrigo Nogueira, Cleber Zanchettin, and Robson Fidalgo. 2024. Predictive authoring for brazilian portuguese augmentative and alternative communication. *Natural Language Processing*, 31(2):535–558.

Jayr Alencar Pereira, David Macêdo, Cleber Zanchettin, Adriano Lorena Inácio de Oliveira, and Robson do Nascimento Fidalgo. 2022. Pictobert: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial, Emmanuelle Esperanca-Rodier, and Benjamin

Lecouteux. 2020. Providing semantic knowledge to a set of pictograms for people with disabilities: a set of links between WordNet and arasaac: Arasaac-WN. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 166–171, Marseille, France. European Language Resources Association.

Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2015. 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*, pages 110–117, Dresden, Germany. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Vincent Vandeghinste, Ineke Schuurman Leen Sevens, and Frank Van Eynde. 2017. Translating text into pictographs. *Natural Language Engineering*, 23(2):217–244.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

A Prompt used for text simplification

```
Convierte las siguientes oraciones en
lectura simple para que puedan ser usadas
en pictogramas. Debes seguir las
instrucciones descritas a continuación:
- La salida debe ser una lista de
oraciones más simples, cortas y fáciles
de leer.
- Las oraciones generadas deben ser
comprensibles para un niño de 7 u 8 años.
- Las palabras utilizadas en las
oraciones deben ser fáciles de entender
para un niño de esa edad.
- La longitud de cada oración no debe
exceder las 10 palabras.
```

Figure 4: Prompt used for automated text simplification.

B Rule-based algorithm for automatic text-to-pictogram generation.

Algorithm 1: Pseudocode for automatic text-to-pictogram translation.

Input: Initial corpus of sentences, ARASAAC pictogram dictionary
Output: Corpus translated into pictogram sequences

```
foreach sentence in corpus do
    Lemmatize the sentence into (word, lemma) pairs;
    Initialize empty list of pictograms;
    foreach (word, lemma) pair do
        if word exists in dictionary then
            Add corresponding pictogram to list;
            Continue to next pair;
        else
            if lemma exists in dictionary then
                Add corresponding pictogram to list;
            else
                Add word information to list (fallback);
            end
        end
    end
    Check for compound pictograms in sentence;
    if found then
        Replace compound pictogram in sentence;
    end
    Generate sequence of pictogram identifiers from list;
    Store original sentence, translation, and metadata;
end
```

C Prompt used for LLama-based Text-to-Concept Generation

D Annotation Interface

E Annotator Analysis

F Examples with high disagreement

Figure 7 presents two representative examples exhibiting high disagreement among human evaluators, illustrating the sources of variability discussed above. In the first example (top), scores range from 1 to 5 (1, 5, 4, 1, 3, 5), reflecting substantial

```
Eres un psicólogo o pedagogo experto en educación inclusiva y comunicación
aumentativa y alternativa (CAA). Tu tarea consiste en identificar los
conceptos clave de una oración ("Input") para facilitar la posterior
selección de pictogramas adecuados. Sigue cuidadosamente estas
instrucciones:
- Extrae los conceptos clave de la oración: pueden ser acciones,
entidades, atributos o relaciones esenciales para el significado general.
- Los conceptos deben ser simples, visuales y adecuados para personas con
dificultades de comunicación o aprendizaje.
- Puedes mantener conceptos compuestos si tienen un significado claro como
unidad (por ejemplo, jugar fútbol), o separar en elementos individuales si
mejora la claridad (jugar, fútbol).
- Si el sujeto está implícito, agrégalo usando el pronombre personal
adecuado según la conjugación verbal (por ejemplo, yo, tú, él).
- Sustituye nombres propios por un término genérico correspondiente como
niño, niña, hombre o mujer, según el género y contexto.
- En preguntas, conserva como concepto clave el pronombre interrogativo
principal (qué, quién, dónde, 33 cuál, etc.).
- Ordena los conceptos en una secuencia que represente claramente el
significado completo de la oración, manteniendo coherencia gramatical y
semántica. La secuencia debe funcionar como una traducción pictográfica.
- Retorna únicamente la lista de conceptos clave, separados por comas. No
incluyas explicaciones, títulos ni ningún otro texto adicional.
```

Figure 5: Prompt used for converting text into a sequence of concepts.

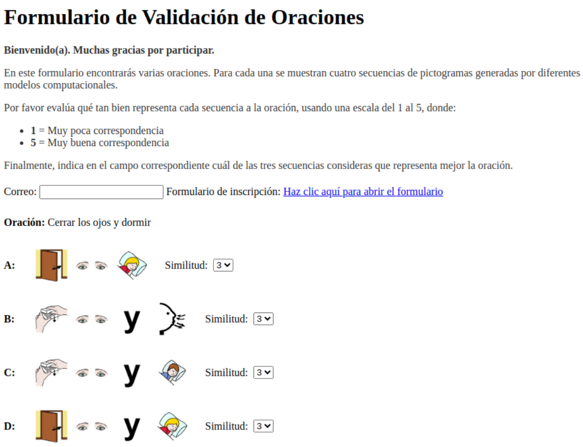

Figure 6: Annotation Interface.

Annotator	LOO-α	Deviation from consensus	Spearman Corr. (mean $\pm$ std)
1	0.208	0.865	0.38 ± 0.28
2	0.194	0.905	0.44 ± 0.25
3	0.185	0.755	0.47 ± 0.26
4	0.284	1.380	0.32 ± 0.28
5	0.206	0.915	0.38 ± 0.27
6	0.187	0.750	0.47 ± 0.24
7	0.258	1.220	0.31 ± 0.29
8	0.187	0.910	0.47 ± 0.25

Table 6: Annotator-level diagnostic analysis based on agreement and deviation metrics.

divergence in judgement. This variation appears to arise from differences in how annotators prioritise semantic components: some strongly penalise inaccuracies in key predicates (e.g., "descubrir" / discover), while others adopt a more tolerant interpretation if the overall scene remains partially interpretable. Additionally, the absence of relational elements (e.g., connectors between the bones and the dinosaur) may hinder compositional understanding, further contributing to inconsistent ratings.

The second example (bottom) shows a similarly wide spread of scores (4, 5, 1, 5, 1, 3), but for a different reason. Here, the intended concept is "inflammation," yet the system selects a pictogram associated with the umbilical cord. Crucially, annotators do not have access to the textual labels underlying ARASAAC symbols and must rely solely on visual interpretation. As a result, some evaluators interpret the pictogram as representing inflammation due to its visual appearance, assigning higher scores despite the semantic mismatch.

These examples highlight two complementary sources of disagreement: (i) differences in evaluative criteria (e.g., strict vs. tolerant semantic assessment), and (ii) intrinsic ambiguity in the pictograms themselves. This supports our earlier findings that disagreement is not solely attributable to annotator noise, but also reflects genuine subjectivity and interpretative variability inherent to the task.

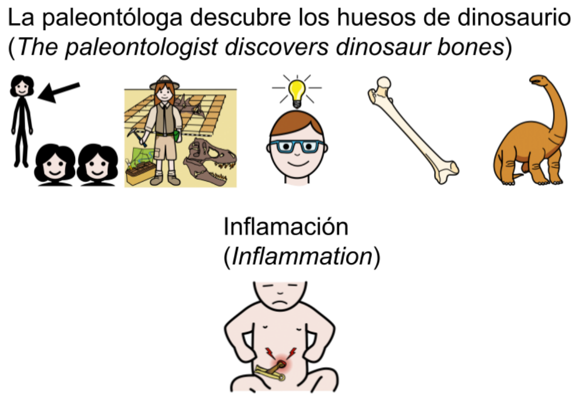

Figure 7: Examples with high disagreement before and after annotator filtering.

G Results before annotator filtering

Table 7 reports the average human evaluation scores for each system. ARAWORD achieves the highest mean score (3.04), confirming its role as a strong baseline. Among neural approaches, T5 benefits substantially from the Text-to-Concept-to-Pictogram setting, reaching a mean score of 2.79 and clearly outperforming the direct Text-to-Pictogram variant (2.18). This improvement indicates that introducing an explicit conceptual layer helps reduce semantic ambiguity in pictogram generation. LLaMA attains a moderate score (2.42), suggesting reasonable performance, albeit below that of the Text-to-Concept-to-Pictogram T5.

Approach	Avg. Score
ARAWORD	**3.04***
T5 (Text-to-Pictogram)	2.18
T5 (Text-to-Concept-to-Pictogram)	**2.79***
LLaMA-8B-1-SHOT	2.42

Table 7: Human evaluation scores before annotator filtering. *Indicates no statistically significant difference.

Statistical analysis confirms that these differences are significant. A Friedman test reveals overall variation across systems ($\chi^2 = 20.68$, $p < 0.05$). Post-hoc Wilcoxon tests ($p < 0.05$) with Holm correction show that ARAWORD significantly outperforms both T5 Text-to-Pictogram and LLaMA, while T5 Text-to-Concept-to-Pictogram yields a significant improvement over its direct counterpart. No significant difference is observed between ARAWORD and the concept-based T5 model.