DUTH at CLEF JOKER 2025 Tasks 2 and 3: Translating Puns and Proper Names with Neural Approaches

Georgios Arampatzis*, Avi Arampatzis

Democritus University of Thrace, Department of Electrical and Computer Engineering, Xanthi, Greece

Abstract

This paper presents the participation of Team DUTH (Democritus University of Thrace) in the CLEF 2025 JOKER shared tasks on computational humor, specifically focusing on the translation of puns (Task 2) and onomastic wordplay (Task 3) from English to French. These tasks pose significant challenges for neural machine translation (NMT) systems due to semantic ambiguity, linguistic creativity, and cultural specificity.

For Task 2, we employed a hybrid architecture that combines multiple NMT systems with a manually curated fallback lexicon. This approach yielded a BLEU score of **41.11** and a BERTScore F_1 of **86.96**, demonstrating improved robustness and humor preservation compared to neural-only baselines.

In Task 3, we addressed the translation of fictional and culturally marked character names using multilingual generative models integrated with rule-based dictionaries and a tiered fallback strategy. Evaluation included both automatic metrics and expert human assessments, revealing key insights into the limits of reference-based evaluation in creative translation tasks.

Our findings suggest that hybrid NMT systems enriched with linguistic insight provide tangible benefits for humor-aware translation. We conclude by outlining directions for future work, including dynamic fallback strategies, humor-sensitive evaluation, and culturally adaptive generation techniques.

Keywords

Wordplay Translation, Humor in NLP, Neural Machine Translation, Fallback Mechanisms, Onomastics

1. Introduction

The translation of wordplay represents a longstanding challenge in both computational linguistics and translation studies. Puns and onomastic humor exploit linguistic ambiguity, phonological similarity, cultural references, and multi-layered meanings—features that often resist straightforward cross-linguistic transfer [1, 2]. Within machine translation (MT), these characteristics demand models with refined semantic, pragmatic, and cultural sensitivity.

The CLEF 2025 JOKER shared tasks serve as a benchmark for evaluating multilingual humor processing systems. This edition comprises two subtasks: Task 2, focusing on the translation of English puns into French, and Task 3, addressing the translation of culturally and phonetically marked character names (onomastic wordplay) [3]. Task 2 emphasizes the preservation of ambiguity and humorous effect, while Task 3 involves creative name generation that resonates across cultural and linguistic boundaries.

As outlined in the official lab overview [3], the 2025 JOKER campaign encourages approaches that move beyond standard translation pipelines and explore linguistically-informed, culturally-aware solutions. It further highlights the inadequacy of conventional evaluation metrics—such as BLEU—for capturing humor-related phenomena, recommending hybrid evaluation schemes that integrate human judgment [4].

In this context, our team from Democritus University of Thrace (DUTH) participated in both tasks of the 2025 edition, investigating hybrid translation workflows that combine multilingual neural machine translation (NMT) models with manually curated fallback strategies. Our previous work on multilingual affective tasks has shown promising results using ensemble methods [5], suggesting that

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

*Corresponding author.

†These authors contributed equally.

geoaramp@ee.duth.gr (G. Arampatzis*); avi@ee.duth.gr (A. Arampatzis)

0009-0003-3840-4537 (G. Arampatzis*); 0000-0003-2415-4592 (A. Arampatzis)

hybrid architectures can improve performance in nuanced language processing. We hypothesize that such hybrid systems can more effectively manage the ambiguity, creativity, and cultural specificity inherent to humor. Our hybrid translation framework aligns with prior approaches that incorporate discrete lexical knowledge into NMT pipelines to better handle rare or culturally marked terms, as in Arthur et al. [6], who used lexical probabilities derived from external resources to guide the translation process.

This work contributes to the growing field of computational humor within NLP, where tasks such as pun translation and humorous name adaptation pose unique challenges. Standard MT systems often fail to preserve semantic content alongside stylistic and cultural nuances, motivating the need for workflows that blend neural generation with human-informed mechanisms.

In Section 2, we describe our hybrid translation framework, outlining the datasets, multilingual models, and fallback mechanisms used for Tasks 2 and 3. Section 3 presents the experimental results obtained through both automatic and manual evaluation metrics, highlighting model performance and limitations. Finally, Section 4 offers concluding remarks and discusses directions for future research in humor-aware machine translation.

2. Approach

2.1. Task 2: Translation of Puns

The translation of puns presents a unique challenge in cross-linguistic humor processing, as it requires not only the preservation of semantic content but also the recreation of linguistic ambiguity, wordplay, or cultural references in the target language. Unlike standard translation tasks, pun translation often involves creative reformulation to maintain both the humorous effect and the intended meaning (Delabastita, 1996). In many cases, literal translation is insufficient, prompting translators—or models—to opt for adaptation strategies such as compensation, modulation, or substitution. Evaluating pun translations thus requires attention to both lexical accuracy and pragmatic impact, particularly when multilingual humor is involved.

2.1.1. Dataset

The dataset used for Task 2 (Wordplay Translation) is composed of a training and a test partition, each exhibiting distinct characteristics. The training set contains 5,838 English–French sentence pairs derived from 1,405 unique English puns. Each pun is accompanied by multiple human-produced translations, providing lexical and stylistic diversity that supports robust model training and evaluation. In contrast, the test set includes 4,537 unique English puns, without any reference translations, and is designed exclusively for blind evaluation. A summary of the dataset composition is provided in Table 1.

Table 1
Number of Entries in the Train and Test Sets (Task 2)

Dataset	Total Entries
Train Set	5838
Test Set	4537

Table 2 presents a comparative overview of the datasets used exclusively for Task 2 . The training set consists of 1,405 unique English puns, each accompanied by multiple French translations, resulting in a total of 5,838 entries. These multiple reference translations provide valuable linguistic diversity and are essential for training and evaluating models on complex tasks such as pun translation.

On the other hand, the test set includes 4,537 unique English puns without any associated reference translations. This reflects the nature of the evaluation phase, where participants' translations are assessed independently against withheld references, either through human judgments or automated evaluation metrics. The test set thus serves as a blind evaluation benchmark.

Table 2
Dataset Statistics for Task 2 (Train and Test Sets)

Dataset	Unique English Puns	Total Entries	Has French Translations
Train Set	1405	5838	Yes
Test Set	4537	4537	No

2.1.2. Models Used

Our translation system incorporated a diverse suite of neural machine translation (NMT) models to ensure stylistic coverage, robustness, and fallback capability. The following models were used:

We included **Google Translate**, a commercial-grade NMT engine commonly used as a general-purpose baseline in multilingual NLP tasks [7]. In addition, we integrated **Argos Translate**, an open-source, offline translation system that supports lightweight and customizable deployment [8].

From the **Helsinki-NLP/OPUS-MT** family, we used the standard `opus-mt-en-fr` model, trained on multilingual OPUS corpora [9]. This model provides robust translation for multiple domains and maintains compatibility with low-resource language scenarios.

We employed **MBART50**, a multilingual encoder-decoder model trained on 50 languages, supporting both supervised and zero-shot translation [10]. We also used **M2M100** in two configurations, including the 1.2B parameter variant, which enables direct source-to-target multilingual translation without pivoting through English [11]. To further explore system diversity, we created a **combined M2M100** configuration that integrates outputs from multiple M2M-based systems.

Additionally, we utilized **T5-base**, a sequence-to-sequence transformer trained under the text-to-text paradigm, enabling translation to be framed as a generative task [12]. In select experiments, we tested its instruction-tuned variant, **FLAN-T5**, as part of a fallback mechanism (*google_flant5_fallback*) to evaluate its responsiveness to stylistic prompts in translation [13].

To explore multilingual generative modeling, we experimented with **BLOOMZ-3B**, a multitask instruction-tuned large language model (LLM) designed for zero- and few-shot transfer across both languages and tasks [14].

Finally, we incorporated a custom **hybrid_fusion** configuration that combines multiple translation engines into a single output through a fusion strategy, aiming to maximize lexical coverage and stylistic adequacy across diverse inputs.

2.1.3. Methodology

To address the challenges of pun translation in Task 2, we adopted a hybrid methodology in which manually curated translations were prioritized, and neural machine translation (NMT) systems served as a fallback mechanism. We integrated a diverse suite of translation engines, including **Google Translate**, **Argos Translate**, the **Helsinki-NLP/opus-mt-en-fr** model [15], Facebook's **M2M100 (418M and 1.2B)** [11], **MBART50** [10], and **T5-base** [12]. We also evaluated its instruction-tuned variant, **FLAN-T5**, as part of a fallback mechanism (*google_flant5_fallback*) [13]. Additionally, we explored two composite configurations: **hybrid_fusion**, which combines multiple systems to maximize lexical coverage and stylistic adequacy, and **combined_m2m100**, which integrates outputs from different M2M-based models.

Each system was embedded within a unified two-stage translation pipeline:

1. A static fallback dictionary containing manually crafted French translations for a selected subset of semantically and stylistically complex English puns.
2. An NMT engine used for all remaining inputs.

All manual translations were created and reviewed by the authors, with emphasis on preserving both semantic fidelity and humorous effect.

At inference time, the system first checked whether the input pun matched an entry in the fallback dictionary. If so, the corresponding manual translation was used directly. Otherwise, the input was

processed by the designated MT model. This approach ensured consistent handling of difficult cases and improved overall robustness.

Our hybrid translation framework aligns with prior approaches that incorporate discrete lexical knowledge into NMT pipelines to better handle rare or culturally marked terms, as in Arthur et al. [6], who used lexical probabilities derived from external resources to guide the translation process.

2.2. Task 3: Onomastic Wordplay Translation

2.2.1. Dataset

The dataset used in Task 3 centers around the translation of fictional, humorous, and culturally marked character names from English to French. It is divided into a training and a test partition. The training set includes 353 entries, each consisting of an English name, a short descriptive context, and a human-produced French translation. These annotated pairs serve as the basis for training or guiding generative systems. The test set comprises 2,696 English entries without accompanying French references, reflecting the task's creative nature and supporting blind evaluation. Table 3 presents the distribution of entries across the two sets.

Table 3
Dataset Statistics for Task 3 (Onomastic Translation)

Dataset	Total Entries
Train Set	353
Test Set	2696

To assess the linguistic and creative complexity of the character names in the Task 3 training set, we conducted a pattern-based classification. Each name was manually or semi-automatically assigned to one of four categories: (i) Alliteration (e.g., names with repeated initial sounds), (ii) Wordplay (e.g., names based on puns, homophones, or semantic ambiguity), (iii) Realistic names with no humorous intent or transformation, and (iv) Unclassified names that either defy clear categorization or require deeper cultural or contextual interpretation. Table 4 presents the approximate distribution of these categories within the training set.

Table 4
Indicative Distribution of Linguistic Patterns in Task 3 Character Names

Pattern Type	Count	% of Train Set
Alliteration	82	23.2%
Wordplay	103	29.2%
Realistic Name	52	14.7%
Unclassified	116	32.9%

2.2.2. Models Used

To support onomastic translation in Task 3, we employed a diverse set of multilingual and generative models with different but complementary capabilities in lexical control, generative capacity, and phonetic variation:

- **Helsinki-NLP/opus-mt-en-fr** and **opus-mt-tc-big-en-fr**: Transformer-based MarianMT models trained on OPUS corpora, offering robust translation across many language pairs [16].
- **facebook/nllb-200-distilled-600M** and **facebook/nllb-200-1.3B**: Large-scale multilingual models trained on over 200 languages, developed by Meta to support low-resource translation via self-supervised dense representation learning [17].

- **facebook/m2m100_418M** and **facebook/m2m100_1.2B**: Multilingual models by Meta AI that support direct translation between 100+ language pairs without pivoting through English, improving semantic alignment across typologically diverse languages [11].
- **T5-base** and **T5-small**: Models from the Text-to-Text Transfer Transformer family [12], treating all NLP tasks as text generation problems, and offering flexible generative capabilities suited for creative name adaptation.
- **MarianMT_BLOOM**: A custom configuration combining the MarianMT architecture with BLOOM embeddings to support high-variance translation behavior.

2.2.3. Methodology

We used the official datasets and evaluation procedures provided by the organizers of Tasks 2 and 3 [18, 19]. Task 3 addressed the challenge of translating fictional character names and stylized proper nouns from English into French, a process demanding not only semantic accuracy but also sensitivity to cultural context and humorous intent. To this end, we developed a hybrid translation framework that integrates several multilingual pretrained neural machine translation (NMT) models with a manually curated dictionary. This approach seeks to combine the broad generalization capacity of large-scale language models with human-informed linguistic and cultural insights.

The translation system utilized a diverse array of pretrained NMT models sourced from the Hugging Face Transformers library. Specifically, we employed the following models: `facebook/m2m100_1.2B` and `facebook/m2m100_418M` [11], `facebook/nllb-200-1.3B` and `facebook/nllb-200-distilled-600M` [17], `t5-small` and `t5-base` [12], `Helsinki-NLP/opus-mt-en-fr` and `Helsinki-NLP/opus-mt-tc-big-en-fr` [9], as well as `MarianMT_BLOOM`. All models were used with their default pretrained weights and standard inference settings.

For models based on the T5 architecture, we adopted task-specific prompting using the format `translate English to French: <name>`. In the case of the M2M100 and NLLB-200 families, we explicitly specified language codes during tokenization (e.g., `eng_Latn` → `fra_Latn`). Translations were generated using greedy decoding via the `model.generate(...)` function, with a maximum token length capped at 128.

The translation pipeline implemented a structured three-stage fallback strategy. First, if the input string matched an entry in a curated bilingual dictionary of proper names, the corresponding French equivalent was directly retrieved. If no match was found, the string was sequentially processed by the NMT models, with their outputs being evaluated for errors, malformations, or runtime failures (e.g., out-of-memory exceptions). When model inference was unsuccessful or produced invalid results, the system defaulted to an external translation service using the Google Translate API, accessed programmatically via the Python library `deep-translator`.

A central component of the pipeline was the manually constructed bilingual dictionary, which included 25 character names selected for their cultural specificity and stylistic complexity. These translations were produced using domain expertise in literary and humorous translation. Where applicable, established canonical French translations were retained (for instance, from *Astérix* or *Charlie et la chocolaterie*). In other cases, novel translations were devised to maintain cultural relevance, phonetic plausibility, and humorous resonance. Illustrative examples include the transformation of `Dogmatix` into `Idéfix`, `Slugworth` into `Espionix`, and `Violet Beauregarde` into `Violetix`.

This design choice aligns with prior research on integrating symbolic lexical knowledge into NMT systems. Arthur et al. [6] demonstrated the utility of using lexical probabilities from external resources to enhance translations involving rare or culturally marked terms. Similarly, our curated dictionary supports translation robustness in cases where standard generative models tend to underperform.

Finally, each translated name was tagged with a binary metadata flag (`manual: 1` or `manual: 0`) to indicate whether it originated from the manual dictionary or from model-based inference. This annotation schema facilitated downstream analysis, particularly in evaluating the performance and reliability of automatic versus human-guided translation pathways.

2.3. Implementation and Environment

All systems were implemented in Python 3.10, using the PyTorch deep learning framework (version 2.7.0) in combination with the Transformers library (version 4.52.4) for model loading, tokenization, and inference. For multilingual translation tasks, we employed pretrained sequence-to-sequence models such as `facebook/m2m100` (418M and 1.2B), `facebook/nllb-200` (1.3B and distilled-600M), `MBART50`, `Helsinki-NLP/opus-mt-en-fr` and `opus-mt-tc-big-en-fr`, `t5-small`, `t5-base`, and `MarianMT_BLOOM`, all accessed through the Hugging Face Transformers API.

Inference was executed on an NVIDIA RTX A6000 GPU with 48 GB of memory, under CUDA 12.5, ensuring fast and stable model performance. All translation outputs were generated in batch mode using efficient GPU memory allocation strategies. A fallback dictionary was incorporated into each system to override specific translation cases with manually curated references.

Complementary translation backends included Argos Translate (via `ctranslate2`) and the Google Translate API, both integrated for comparative evaluation.

3. Results

3.1. Evaluation Metrics

For Task 2, system outputs are initially ranked using the BLEU metric [20], which computes n-gram overlap between the machine-generated translation and one or more human references. While BLEU remains a standard benchmark for translation quality, it has well-known limitations when applied to pun translation, as it does not account for semantic adequacy, humor preservation, or stylistic variation.

To address this, the final evaluation will be performed manually by expert annotators. Human judges will assess each translation based on criteria including semantic field preservation, sense equivalence, syntactic and lexical fluency, and the degree to which the original wordplay is retained or effectively adapted in French. This hybrid evaluation procedure ensures that systems are not only rewarded for lexical similarity, but also for their ability to capture the linguistic creativity inherent in puns.

For Task 3, the official evaluation procedure is based on *exact match accuracy* between the submitted French names and a set of manually curated reference translations. This automatic scoring rewards systems that generate character names identical to one of the provided references, thus emphasizing lexical precision.

Nevertheless, given the creative and culturally embedded nature of onomastic wordplay, a final manual evaluation will be conducted by expert annotators. The human evaluation will consider several qualitative aspects, including preservation of the semantic field, cultural appropriateness, linguistic creativity, and the extent to which the translated name preserves or effectively transforms the original wordplay. This two-stage evaluation protocol ensures a balanced assessment that goes beyond surface-level lexical overlap.In addition to BLEU, we employed complementary automatic metrics to obtain a broader evaluation perspective. BERTScore [21] was used to assess contextual semantic similarity between system outputs and references, providing finer-grained signals than surface n-gram overlap. For Task 2, we also evaluated pun location detection accuracy, measuring each model's ability to correctly identify the position of wordplay within the sentence.

These additional metrics allowed us to triangulate translation performance from both lexical and semantic standpoints. Combined with the expert-driven manual evaluation, this multi-layered framework ensured a more comprehensive and nuanced assessment of systems' ability to preserve meaning, humor, and creativity in multilingual pun translation.

3.2. Task 2: Translation of Puns

3.2.1. Experimental Results

Table 5 presents the BLEU scores and n-gram precision metrics for the systems submitted by team DUTH in Task 2, sorted by BLEU score. The top-performing model was `hybrid_fusion` (**BLEU = 41.11**),

followed closely by helsinki (BLEU = 41.01). Both exhibited high unigram and bigram precision, indicating strong lexical overlap with the reference translations.

Close behind were several Google Translate variants—including GoogleTranslate_fallback and google_flanT5_fallback—which consistently achieved BLEU scores above 40.7 and maintained solid performance across n-gram orders up to $n = 4$. This suggests their effectiveness in capturing both surface-level and moderately complex lexical patterns.

argos and m2m100_1_2B also performed competitively, with BLEU scores of 40.49 and 36.46, respectively. Transformer-based models such as mbart50 and t5_base showed more modest performance (BLEU 32–33), while combined_m2m100 scored 30.00.

At the lower end, bloomz3b scored significantly lower (BLEU = 16.68), reflecting limited transfer capabilities in low-resource or stylistically marked inputs.

Overall, these results confirm that pretrained multilingual NMT models—especially those enhanced via fallback mechanisms—remain strong baselines for pun translation. Incorporating hybrid or ensemble methods, as seen in hybrid_fusion, appears to offer further gains in robustness and lexical coverage.

Table 5
BLEU score and n-gram precision for the models submitted by team DUTH, sorted in descending order of BLEU score. Results from [3].

Model	BLEU	$n = 1$	$n = 2$	$n = 3$	$n = 4$
hybrid_fusion	41.11	63.45	44.62	35.17	28.70
helsinki	41.01	63.40	44.52	35.07	28.58
GoogleTranslate_fallback	40.94	62.75	44.21	35.12	28.84
google_flant5_fallback	40.74	62.60	43.99	34.92	28.65
GoogleTranslate	40.73	62.59	43.98	34.91	28.64
GoogleTranslate_fallback	40.73	62.60	43.99	34.91	28.63
argos	40.49	63.21	44.13	34.73	28.24
m2m100_1_2B	36.46	61.22	41.01	30.91	23.90
mbart50	32.73	57.21	36.32	26.81	20.62
t5_base	32.44	56.04	36.26	26.82	20.33
combined_m2m100	30.00	56.97	34.81	24.11	17.66
bloomz3b	16.68	41.17	19.57	12.15	7.91

Table 6 presents the BERTScore results (Precision, Recall, and F_1) for the 12 systems submitted by team DUTH for Task 2.

The highest F_1 scores were obtained by GoogleTranslate_fallback and GoogleTranslate, both achieving **86.96**, demonstrating the semantic robustness of large pretrained translation engines when enhanced with fallback mechanisms. Close behind was google_flant5_fallback with an F_1 of **86.93**, confirming the effectiveness of instruction-tuned Transformer models in preserving meaning.

Other systems—such as hybrid_fusion and argos—also performed competitively, scoring above 86.4 in F_1. These results underscore the value of hybrid strategies in semantically rich and stylistically demanding translation tasks, especially in creative linguistic settings.

The middle-performing group includes helsinki (**86.40**) and m2m100_1_2B (**85.56**), indicating that multilingual models and traditional NMT approaches can still deliver strong results with proper adaptation.

In contrast, models like t5_base, and bloomz3b scored below 84 in F_1, suggesting weaker semantic alignment. This may be attributed to limited fine-tuning or inadequate adaptation to the task's stylistic demands.

Beyond the F_1 score, precision and recall provide complementary insights into system behavior. Top-performing systems combined high precision and recall, demonstrating both lexical accuracy and contextual alignment. In contrast, weaker systems struggled with semantic fidelity, particularly in humorous or culturally marked content. These findings reaffirm the importance of fallback-enhanced and hybrid systems for producing semantically faithful translations in creative, multilingual contexts.

Table 6
BERTScore results (Precision, Recall, F_1) for the 12 models submitted by team DUTH, sorted in descending order by F_1. Source: [3].

Model	P	R	F_1
GoogleTranslate_fallback	87.20	86.77	86.96
GoogleTranslate	87.20	86.77	86.96
google_flant5_fallback	87.17	86.74	86.93
hybrid_fusion	87.18	85.79	86.43
argos	87.00	85.91	86.42
helsinki	87.17	85.74	86.40
GoogleTranslate_fallback	86.34	85.91	86.10
m2m100_1_2B	85.90	85.28	85.56
mbart50	85.14	84.14	84.60
combined_m2m100	84.76	84.05	84.37
t5_base	83.91	83.60	83.71
bloomz3b	79.30	78.66	78.94

Table 7 presents the evaluation results for twelve models submitted by team DUTH (Democritus University of Thrace), focusing on their ability to correctly identify the location of puns within sentence-level contexts. The evaluation was conducted on a dataset comprising **1682 instances**, corresponding to the total number of annotated pun occurrences assigned to each model for assessment.

The table includes four columns: the model name; the *Total* column, indicating the number of examples evaluated (uniformly 1682); the *Correct* column, listing the number of correctly identified pun locations; and the *(%)* column, which expresses accuracy as a percentage of the total, calculated as:

$$\text{Accuracy (\%)} = \frac{\text{Correct}}{1682} \times 100.$$

The results are sorted in descending order of location accuracy. The highest-performing model, `helsinki`, achieved **6.72%** accuracy (113 correct identifications). It was followed by `GoogleTranslate_fallback`, `hybrid_fusion`, and `google_flant5_fallback`, each reaching **6.66%** (112 correct), while `GoogleTranslate` obtained **6.60%** and `argos` scored **6.48%**.

Mid-performing systems included `t5_base` (**5.65%**) and `m2m100_1_2B` (**5.41%**), whereas `mbart50` and `combined_m2m100` achieved **4.82%** and **4.22%**, respectively. The lowest performance was by `bloomz3b`, at **2.68%** (45 correct). These results underscore a gap between top-ranked systems and multilingual or instruction-tuned models that lack task-specific adaptation.

Table 7
Pun location detection accuracy for models submitted by team DUTH, based on 1682 instances. Sorted by descending percentage. Results from [3].

Model	Total	Correct	(%)
helsinki	1682	113	6.72
GoogleTranslate_fallback	1682	112	6.66
hybrid_fusion	1682	112	6.66
google_flant5_fallback	1682	112	6.66
GoogleTranslate	1682	111	6.60
GoogleTranslate_fallback	1682	111	6.60
argos	1682	109	6.48
t5_base	1682	95	5.65
m2m100_1_2B	1682	91	5.41
mbart50	1682	81	4.82
combined_m2m100	1682	71	4.22
bloomz3b	1682	45	2.68

3.3. Task 3: Onomastic Wordplay Translation

3.3.1. Experimental Results

Table 8 presents the evaluation results of twelve machine translation systems for Task 3, assessed across three metrics: *Automatic*, *Manual*, and *Identical*.

The *Automatic* score indicates the percentage of lexical matches with the reference translations, based on string-level comparisons. While informative, it does not guarantee semantic adequacy. The *Manual* score reflects the proportion of outputs deemed acceptable by human evaluators, taking into account meaning preservation, discourse fluency, and cultural appropriateness. The *Identical* metric captures the percentage of outputs that are exact copies of the source input—typically signaling failure to translate rather than successful output generation.

The results reveal substantial variation in system performance. The system `Helsinki` achieved the highest automatic score (14.83%) and shared the highest manual score (18.88%), yet also exhibited a 77.67% identical rate. This suggests that a large portion of its outputs were simple copies of the input, which happened to match the references, artificially inflating its automatic score without necessarily indicating high translation quality.

Several systems, such as the `facebook-nllb-200` variants and `MarianMT_BLOOM`, displayed more balanced behavior across all three metrics (10–11% *Automatic*, 13–16% *Manual*, around 45% *Identical*). These results point to partial translation adequacy with a conservative output strategy that favors safe but modest transformations.

In contrast, low-performing models like `facebook-m2m100_1.2B` and `facebook-m2m100_418M` showed weak performance across all metrics. Their low *Identical* scores (below 21%) suggest that failure stemmed not from excessive copying, but from an inability to generate semantically valid translations.

At the other extreme, one system from the `Helsinki` family exhibited a 100% identical rate alongside a manual score of just 2.55%. Although it achieved a non-zero automatic score (11.83%), this clearly reflects a complete failure in semantic transformation, highlighting the limitations of automatic metrics when lexical overlap lacks meaningful equivalence.

Overall, the findings underscore that no single metric suffices for evaluating performance on linguistically creative tasks such as pun translation. A comprehensive approach that combines automatic scoring, human judgment, and behavioral indicators like source copying is essential for reliably assessing translation quality in semantically complex and stylistically sensitive contexts.

Table 8
Evaluation results showing the percentage of matching translation instances for each model, based on automatic and manual evaluation, as well as the proportion of translations identical to the source text. Results from [3].

Model	Automatic	Manual	Identical
Helsinki	14.83	18.88	77.67
Helsinki-NLP-opus-mt-tc-big-en-fr	14.66	18.88	77.45
Helsinki	11.83	2.55	100.00
facebook-nllb-200-distilled-600M	10.72	16.75	41.83
facebook-nllb-200-1.3B	10.72	16.75	41.83
MarianMT_BLOOM	10.42	13.86	45.95
MarianMT_BLOOM	10.29	13.86	45.78
Helsinki-NLP-opus-mt-en-fr	10.29	13.86	45.78
t5-base	8.57	7.03	50.32
t5-small	8.53	6.00	58.04
facebook-m2m100_1.2B	4.71	9.50	19.12
facebook-m2m100_418M	4.37	4.00	20.15

The results reveal several important observations:

One system exhibits a 100% identical rate with almost negligible performance in the manual evaluation (2.55%). This outcome strongly suggests a complete failure in the translation process, despite registering

a non-zero automatic score (11.83%). This highlights the **misleading nature of automatic scores** when systems achieve lexical overlap without true semantic transformation.

Conversely, another system reaches a 77.45% identical rate while achieving the **highest automatic score (14.66%)**. This suggests that much of the copied content happened to coincide with the reference output, potentially **inflating the automatic evaluation** and masking translation inadequacy when judged by humans.

Several systems demonstrate moderate values across all three metrics—with automatic scores between 10–11%, manual scores between 13–14%, and identical outputs around 45%. This distribution implies partial translation adequacy accompanied by a conservative output strategy, where systems attempt meaningful translation but lean toward lexical safety.

Other systems show consistently low performance across all metrics (e.g., scores below 10% in both automatic and manual evaluations) combined with low identical rates. This pattern suggests not copying, but rather a limited capacity to generate semantically valid translations.

In summary, the results reveal substantial variation in system behavior, ranging from source copying to partial adequacy to complete failure. It becomes evident that **no single metric is sufficient** for evaluating performance in translation tasks that involve linguistic creativity. A comprehensive evaluation approach—one that integrates automatic scores, human judgments, and behavioral indicators such as copying—offers a more reliable and multidimensional assessment of translation quality in contexts where semantic ambiguity, creativity, and nuanced expression play a central role.

4. Conclusion and Future Work

Building upon insights from previous CLEF JOKER evaluations [22, 23], this study advances the exploration of hybrid methods for humor-aware machine translation. Through our participation in Tasks 2 and 3 of the 2025 edition, we demonstrated the potential of fallback-enhanced systems to manage ambiguity, cultural specificity, and linguistic creativity in multilingual contexts.

This paper explored hybrid approaches to humor-aware translation through participation in the CLEF 2025 shared tasks JOKER. Task 2 focused on the translation of puns from English into French, while Task 3 addressed the creative rendering of culturally marked fictional character names. In both tasks, our hybrid system—combining neural machine translation (NMT) models with manually curated fallback mechanisms—demonstrated enhanced robustness and stylistic fidelity compared to purely neural systems.

In Task 2, hybrid systems consistently outperformed neural baselines in BLEU and BERTScore metrics and received favorable human judgments for preserving humor and meaning. Nevertheless, the divergence between automatic and manual evaluations confirmed the inadequacy of surface-level metrics in humor translation. In Task 3, our manually constructed bilingual dictionary proved essential for handling phonologically and culturally complex names, particularly when established or contextually appropriate translations were required.

Future work will proceed along several directions. First, we plan to expand our dictionaries through corpus mining and crowd-sourcing techniques, especially for culturally rich entries. Second, we aim to develop classifiers capable of detecting linguistic patterns in character names (e.g., alliteration, puns, cultural references) to dynamically guide translation strategies. Third, we will experiment with prompt engineering and stylistic control techniques in instruction-tuned models (e.g., FLAN-T5, mT5, BLOOMZ) to enable zero- and few-shot name translation capabilities.

This study contributes to the emerging field of computational humor and highlights the importance of integrating symbolic knowledge, linguistic expertise, and creative reasoning in multilingual NLP systems.

5. Acknowledgements

We thank the organizers of the CLEF 2025 JOKER shared task for providing high-quality data, tools, and a clear, well-designed evaluation framework for computational humor detection. Their contribution to promoting reproducible experimentation, robust comparative evaluation, and scientific progress in the rapidly evolving field of computational humor is greatly appreciated.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Delabastita, Introduction, in: Wordplay and Translation: Essays on Punning and Translation, St. Jerome Publishing, 1996, pp. 1–22.
[2] D. Chiaro, Translation and humour, humour and translation, in: D. Chiaro (Ed.), Translation, Humour and Literature, Continuum, 2010, pp. 1–29.
[3] L. Ermakova, R. Campos, A. Bosser, T. Miller, Overview of the clef 2025 joker lab: Humour in machine, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025. URL: https://www.joker-project.com/, to appear.
[4] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[5] G. Arampatzis, A. Arampatzis, Duth at semeval-2023 task 2: Multilingual complex named entity recognition with cross-lingual ensemble learning, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1238–1243.
[6] P. Arthur, G. Neubig, S. Nakamura, Incorporating discrete translation lexicons into neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 1557–1567.
[7] Google, Google translate, https://translate.google.com, 2025. Accessed July 2025.
[8] A. O. Technologies, Argos translate, https://www.argosopentech.com/, 2025. Version 1.8, Accessed July 2025.
[9] J. Tiedemann, S. Thottingal, Opus-mt – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020.
[10] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, in: Transactions of the Association for Computational Linguistics (TACL), 2020.
[11] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, et al., Beyond english-centric multilingual machine translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.
[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.
[13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[14] N. Muennighoff, et al., Crosslingual generalization through multitask finetuning, arXiv preprint arXiv:2211.01786 (2022).
[15] J. Tiedemann, S. Thottingal, Opus-mt: Building open translation services for the world, arXiv preprint arXiv:2005.11867 (2020).

[16] J. Tiedemann, S. Thottingal, Opus-mt — building open translation services using neural machine translation, https://github.com/Helsinki-NLP/OPUS-MT, 2020. Accessed: 2025-07-03.
[17] M. R. Costa-jussà, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).
[18] L. Ermakova, et al., Overview of the CLEF 2025 JOKER task 2: Wordplay translation from english into french, in: Working Notes of CLEF 2025 Labs and Workshops, CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[19] L. Ermakova, et al., Overview of the CLEF 2025 JOKER task 3: Onomastic wordplay translation, in: Working Notes of CLEF 2025 Labs and Workshops, CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[20] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[22] L. Ermakova, et al., Overview of the JOKER track at CLEF 2023: Automatic wordplay analysis, in: CLEF 2023 Working Notes, CEUR-WS.org, 2023.
[23] L. Ermakova, T. Poibeau, et al., JOKER@CLEF 2024: Challenges in cross-lingual humor translation, in: CLEF 2024 Working Notes, CEUR-WS.org, 2024.