Overview of the CLEF 2026 JOKER Task 2: Pun Translation from English to French

Liana Ermakova[1], Yaël Naud[1], Lilian Binet[1], Baptiste Dumas[1] and Jaap Kamps[2]

[1]HCTI, University of Brest, France
[2]University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper presents an overview of Task 2 of the CLEF 2026 JOKER track, which focuses on translating puns from English into French. The task addresses one of the most challenging problems in machine translation: preserving both the meaning and the humorous wordplay of the source text across languages. Building on previous editions of JOKER, the 2026 task introduces a new manually translated test collection of unseen English puns and evaluates systems using both automatic metrics and expert human assessment. We describe the construction of the benchmark, the evaluation methodology, the participating teams, and the approaches they employed. The submitted systems demonstrate substantial improvements, yet pun translation remains far from solved, as models frequently preserve the literal meaning while failing to recreate the underlying wordplay. This year, we evaluated the submitted runs using the pun location-based metric, which has been shown to correlate strongly with human judgments of successful translations, i.e., translations that both preserve the meaning of the source pun and successfully recreate the wordplay. Our analysis highlights the strengths and limitations of current approaches, identifies the linguistic phenomena that remain most difficult to translate, and discusses the implications for future research on computational humor, creative language generation, and machine translation.

Keywords

Wordplay, Pun, Machine translation, Humour, Large Language Models, LLM

1. Introduction

Humor remains one of the most challenging aspects of natural language processing because it often relies on ambiguity, cultural knowledge, and creative use of language. Among the different forms of humor, puns are particularly difficult to translate, as an effective translation must preserve not only the meaning of the original text but also the underlying wordplay, which frequently depends on language-specific phonetic, lexical, or semantic relationships. Consequently, automatic pun translation represents a demanding benchmark for evaluating the creative and multilingual capabilities of modern machine translation systems and large language models. This paper presents an overview of Task 2 of the CLEF 2026 JOKER Lab [1], which focuses on the automatic translation of English puns into French. For details on other JOKER-2026 tasks, we refer the reader to their respective overview papers: Task 1 (*Humour-aware Information Retrieval*): Retrieve short humorous texts for a query in English and Hinglish [2], Task 3 (*Onomastic Wordplay Translation*): translate onomastic wordplay from English to French [3], and Task 4 (*Humour Generation*): generate humorous text from a given context in English, French, and Spanish [4].

The goal of this task is to translate English punning jokes into French. The translations should preserve, as much as possible, both the meaning and the wordplay of the original joke. This is a particularly challenging task for both modern machine translation systems and professional human translators, as the linguistic mechanisms underlying puns are often language-specific. At the same time, its creative nature makes the task especially engaging for translators, enabling us to involve both translation students and professional translators in creating a high-quality benchmark corpus.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
liana.ermakova@univ-brest.fr (L. Ermakova); kamps@uva.nl (J. Kamps)
https://www.joker-project.com (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

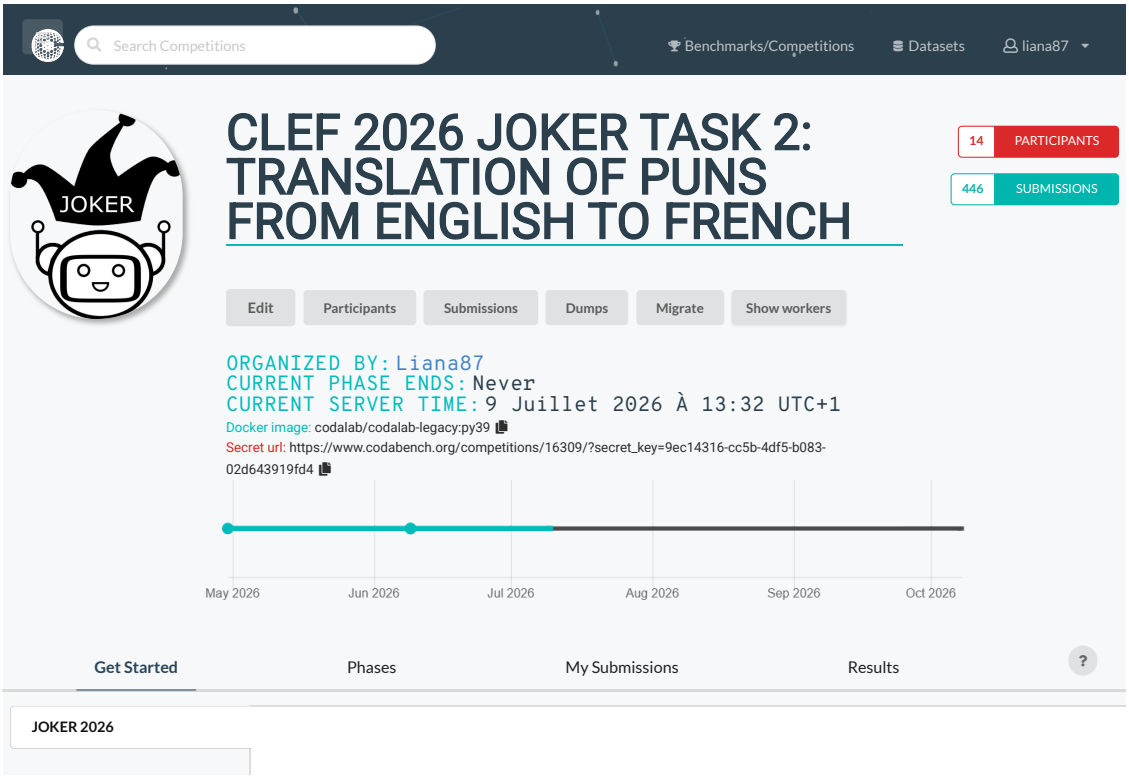

Figure 1: CodaBench for Task 2

Pun translation from English into French has been part of the JOKER lab since its first edition [5, 6], reflecting the longstanding challenge of preserving both meaning and humor across languages. This effort has resulted in the construction of a unique large-scale parallel corpus for wordplay translation [7]. Since the first edition of the task in 2022 [5, 6], the corpus has nearly tripled in size, making it one of the largest publicly available resources for the evaluation of automatic pun translation. Besides, we annotated all translations with pun location, i.e., the word or phrase carrying the double meaning (see examples in Table 1). This pun location-and-interpretation task in English, French, and Spanish was introduced at the JOKER Lab in 2023 [8, 9]. In 2023, we also ran a pilot of the English-to-Spanish pun translation [8, 10]. In 2025, we transitioned to CodaBench[1] for submission management [11, 12]. Although the official ranking has always been based on expert human evaluation, the CodaBench leaderboard in 2025 relied on the BLEU score, which is known to be poorly suited for evaluating pun translation [7, 12]. Before 2025, most participants submitted literal translations, which were successful mainly when the pun relied on cognates shared between English and French. In 2025, however, participant approaches became considerably more sophisticated. For example, the DSGT system used a multi-agent translation framework combining contrastive learning with phonetic-semantic embeddings and was trained on the JOKER corpus [13]. Unlike earlier systems, DSGT frequently generated non-literal translations that recreated the wordplay rather than translating it directly, with approximately 77% of its translations judged valid. In 2025, we also experimented with a pun-location metric, which evaluates whether the translated text contains a word or phrase that carries a double meaning. This metric showed a high correlation with human judgments [12]. Thus, the 2026 edition adopts the pun-location metric as the automatic evaluation measure on CodaBench, while expert manual evaluation remains the additional assessment methodology.

This year, 14 teams registered at CodaBench[2] (see Figure 1) and submitted 446 runs.

The remainder of this paper is organized as follows. Section 2 describes the construction, format, and statistics of the training and test collections. Section 3 presents the evaluation framework, including both automatic and manual evaluation methodologies. Section 3.1 summarizes the participating teams and their approaches. Section 3.2 reports and discusses the experimental results. Finally, Section 4 concludes the paper and outlines future research directions.

2. Data

The training data for Task 2 includes 1,405 instances of wordplay in English and 5,838 professional human translations into French. The test data includes 2,656 English instances and 6,951 reference translations into French by professional translators or valid machine translations from the previous editions of JOKER. All French translations containing wordplay are manually annotated with the words or phrases carrying multiple meanings, i.e. wordplay location. The examples of wordplay location are given in Table 1.

We provide input train and test data in the format of JSON files with the following fields:

- **id_en**: a unique identifier from the input file.
- **text_en**: the text of the instance of source wordplay in English. Note that the English texts are not unique in the file, as the same English pun might have multiple French translations.

Example of an input file:

```
[
    {
        "id_en":"en_1007",
        "text_en":"Save the whales, spouted Tom."
    }
]
```

For training, we also provided a qrel JSON file with the following fields:

- **id_en**: a unique identifier from the input file. Note that this identifier is not unique in the qrel file, as the same English pun might have multiple French translations.
- **text_fr**: translation of the wordplay into French

An example of a qrel file is as follows:

```
[
    {
        "id_en":"en_1007",
        "text_fr":"\"Il faut sauver les baleines\", jeta Tom avant de se
        ↪  tasser."
    },
    {
        "id_en":"en_1007",
        "text_fr":"\"Il faut sauver les baleines\", interjeta Tom."
    },
    {
        "id_en":"en_1007",
        "text_fr":"Moi je sauve les baleines, Tom s'en venta."
```

[1]https://www.codabench.org/competitions/8748/
[2]https://www.codabench.org/competitions/16309/

```
    },
    {
        "id_en":"en_1007",
        "text_fr":"Louis évent-a le projet de sauvetage des baleines."
    },
    {
        "id_en":"en_1007",
        "text_fr":"\"Sauvez les baleines\", proclama Tom à tout évent."
    },
    {
        "id_en":"en_1007","text_fr":"\"Sauvez les baleines, cracha Toto,
        ↪  Cétacé!\""
    }
]
```

The test output was requested to be provided in JSON format with the following fields:

- **run_id**: Run ID starting with <team_id>_<task_id>_<method_used>, e.g. UBO_task_2_Gemini
- **manual**: Whether the run is manual {0,1}
- **id_en**: a unique identifier from the input file
- **en**: the text of the instance of source wordplay in English
- **fr**: translation of the wordplay into French

An ouput example is as follows:

```
[
    {
        "run_id":"UBO_task_2_Gemini",
        "manual":0,
        "id_en":"6",
        "en":"\"Are you drinking again?\"\n\"No, it's just tea.\"\n\"What
        ↪  kind of tea?\"\n\"Tea-quila.\"",
        "fr":"\" Tu t'es remis à boire ? \"\n\" Non, c'est juste du thé.
        ↪  \"\n\" Quelle sorte de thé ? \"\n\" Du thé-quila. \""
    }
]
```

3. Evaluation

In the previous editions of JOKER, we manually evaluated translations in terms of meaning preservation and the presence of wordplay [11, 12]. We additionally provided BLEU and BERTScore. However, these metrics are not aligned with human evaluation [12]. Pun location-based evaluation allows for a more fine-grained analysis of generated translations. We identified words or phrases with multiple meanings (pun locations) in the reference texts by combining French reference translations with pun location annotations. In the previous edition of JOKER, we showed that the location-based metric correlates better with human judgments than traditional metrics [12, 11]. We report the proportion of candidate translations that contain at least one annotated pun-location token among the corresponding reference sentences. Table 1 shows an example of reference translations for an English pun *An electric company is always looking for high **energy** employees* along with their location annotations. A candidate translation of this English pun is considered correct if it contains one of the annotated locations that realizes a double meaning in words related to electricity or energy *{'énergie', 'faisait du remplissage', 'chargés à bloc', 'foudres de guerre', 'gonflés à bloc', 'statiques', 'survoltés', 'pile électrique', 'énergique', 'énergiques', 'surpuissants', 'au courant', 'le courant passe', 'énergétiques', 'branchés'}*. We report the results of the location-based evaluation on both the test and the training sets.

Table 1
Examples of location annotation

English pun / References + Location
An electric company is always looking for high energy employees.
Une compagnie d'électricité recherche toujours des employés débordant d'énergie.
Une compagnie d'électricité cherchent des employés qui ont de l'énergie.,faisait du remplissage
Un fournisseur d'électricité attend de ses employés qu'ils débordent d'énergie.
Une compagnie d'électricité est toujours à la recherche d'employés plein d'énergie.
Les compagnies d'électricité sont toujours à la recherche d'employés plein d'énergie.
EDF cherche en permanence des employés qui ont beaucoup d'énergie.
Un fournisseur d'électricité doit recruter des employés plein d'énergie.
Une société électrique recherche toujours des employés chargés à bloc.
Les compagnies d'électricité cherche toujours des employés qui soient des foudres de guerre.
Les concessionnaires recherchent des employés gonflés à bloc.
Un fournisseur d'électricité recherche toujours des employés qui ont de l'énergie à revendre.
Les compagnies d'électricité recherchent toujours des employés avec beaucoup d'énergie.
Les compagnies d'électricité n'embauchent pas d'employés statiques.
Une compagnie d'électricité est toujours à la recherche d'employés survoltés.
Les employés d'EDF sont comme des piles électriques !
Une entreprise de production d'électricité est toujours à la recherche d'employés énergiques.
EDF est toujours à la recherche d'employés énergiques.
"Dans le domaine de l'énergie, on recherche des employés surpuissants."
Un fournisseur d'électricité est toujours à la recherche d'employés énergiques.
"Dans les centrales électriques, les gens sont toujours au courant."
Les fournisseurs d'électricité cherchent toujours des employés énergiques.
Distributeur d'électricité cherche employés au courant.
Une entreprise d'électricité cherche toujours à recruter des employés énergiques.
Une compagnie d'électricité cherche toujours des employés survoltés.
les compagnies électriques sont toujours à la recherche de gens avec qui le courant passe.
Une compagnie d'électricité cherche toujours des employés qui ont beaucoup d'énergie.
Une compagnie d'électricité est toujours à la recherche d'employés hautement énergétiques.
Une entreprise d'électricité est toujours à la recherche d'employés de haute énergie.
Une entreprise électrique cherche toujours des employés à haute énergie.
Une entreprise électrique est toujours à la recherche d'employés à haute énergie.
Une entreprise électrique est toujours à la recherche d'employés énergétiques.
Une compagnie d'électricité cherche toujours à recruter des employés bien branchés.
Une compagnie d'électricité est toujours à la recherche d'employés chargés à bloc.

We continued the practice of manually evaluating translations with respect to wordplay preservation and meaning preservation [11, 12, 10]. We manually annotated 948 French translations of 100 English puns. The translations were sampled from the top-3 runs per team, ranked by the location-based metric, after removing duplicate and untranslated outputs. The annotation was carried out by three native French-speaking Master's students in translation at the University of Brest. Each translation was initially evaluated by a single annotator. Translations that the first annotator identified as containing an attempt at wordplay were subsequently reviewed by a second annotator, whose judgment was retained for these more challenging cases. We report the manual evaluation only on the test data. Runs are ranked according to the number of successful translations – i.e., translations preserving, to the extent possible, both the form and sense of the original wordplay.

In addition to the official evaluation, we report BLEU scores [14] for both the test and training sets. While BLEU measures lexical overlap with the reference translations, it does not account for the preservation or recreation of wordplay and should therefore be interpreted as a complementary evaluation metric. We used the SacreBLEU implementation [15].

Table 2
Top-5 results per team for Task 2 Pun Translation (pun location metric). # runs is the number of runs submitted by each team.

#	team	# runs	Score	Rank	run_id
1	dsgt-arc	37	32.59	1	ensemble_0_0_0_100_gpt_pooled_transl
2	dsgt-arc	37	32.16	2	gpt_5.5_single
3	dsgt-arc	37	31.11	3	ensemble_0_0_0_100_all_pooled_transl
4	dsgt-arc	37	29.74	4	ensemble_0_0_0_100_claude_pooled_transl
5	dsgt-arc	37	28.88	5	ensemble_0_0_0_100_gemini_pooled_transl
6	n0shi	5	28.71	6	Satir_CVO_Direct_mistral-large:123b
7	arampageos	340	27.94	7	QANCHOR_TOP25
8	arampageos	340	27.90	8	NOTORCH_SELECTOR_STRICT
9	arampageos	340	27.90	9	QANCHOR_REFONLY
10	arampageos	340	27.90	10	NOTORCH_SELECTOR_MICRO5
11	arampageos	340	27.90	11	GROQ_FULL_PARTIAL_5
12	n0shi	5	25.76	325	Satir_Baseline
13	ezrasmink	5	25.22	336	UAms_marian-ft_content0.0
14	ezrasmink	5	25.22	337	UAms_marian-ft_content0.25
15	n0shi	5	25.11	342	Satir_BaselineCVO
16	n0shi	5	24.83	348	Satir_Classifier
17	robinflier	2	24.57	351	UAms_MarianMT_XLMRoBERTa
18	ezrasmink	5	24.47	355	UAms_opus-mt-en-fr-finetuned
19	robinflier	2	23.85	359	UAms_MarianMT_CamemBERT
20	ezrasmink	5	23.58	363	UAms_nllb-ft_content0.25
21	n0shi	5	22.70	368	Satir_ClassifierCVO
22	madhangi	1	22.29	370	PixelPair_groq_llama
23	ezrasmink	5	17.34	384	UAms_mbart-ft_content0.25
24	**Baseline**	1	7.05	391	identity_run

3.1. Participants' approaches

14 registered teams and submitted 446 to CodaBench. We excluded invalid runs and runs with duplicated scores and run identifiers.

Team *dsgt-arc* [16] submitted 37 runs. Their approach is based on a retrieval system that searches semantic and phonological neighborhoods for target-language affordances, various language models for generating translation candidates, and a multi-perspective candidate ranking.

Team *ezrasmink* and *robinflier* (UAmsterdam) [17] submitted 7 runs to Task 2. They fine-tuned MarianMT on the JOKER train data for the pun translation task and then used pun detection classifiers to select pun-preserving translations from multiple candidates. For classification, they used XLM-RoBERTa-base and CamemBERT-base, trained on the JOKER 2023 pun detection data [18].

Team *n0shi* [19] submitted 5 runs to Task 2. They used a retrieval-augmented, multi-candidate generate-and-judge pipeline based on the Constant-Variable Optimization (CVO) prompting strategy from the Pun2Pun benchmark. They further extended this approach with a Delabastita-based pun-type classifier that routes each joke to a type-specific decomposition prompt.

Team *arampageos* [20] submitted 340 runs to Task 2 based on a conservative hybrid strategy. They generated multiple candidate translations using several machine translation systems (e.g., Google Translate, Helsinki-NLP/MarianMT, LLMs) and applied instruction-tuned LLMs (Gemma-2-9B-Instruct and Mistral-7B-Instruct) to perform block-level fusion based on meaning preservation, fluency, and preservation of potential wordplay.

Team *madhangi* (no paper) submitted 1 run to Task 2 without a paper.

Table 3
Top-3 results per team for Task 2 Pun Translation (manual evaluation). WP = percentage of translations containing wordplay; MP = percentage preserving the meaning of the source English pun; Success = percentage preserving both meaning and wordplay. # runs is the number of runs submitted by each team.

#	team	# runs	count	WP	MP	success	Rank	run_id
1	dsgt-arc	37	96	82	48	44	1	ensemble_0_0_0_100_all_pooled_transl
2	dsgt-arc	37	96	79	50	44	2	ensemble_0_0_0_100_gpt_pooled_transl
3	dsgt-arc	37	96	79	50	42	3	gpt_5.5_single
4	n0shi	5	100	38	70	29	13	Satir_CVO_Direct_mistral-large:123b
5	n0shi	5	85	39	60	29	16	Satir_Baseline
6	arampageos	340	97	29	77	25	23	TrainQrelsLeak_FUZZY96_LOCKED_MEDOID
7	arampageos	340	97	29	77	25	24	TrainQrelsLeak_EXACT_ALL_MEDOID
8	arampageos	340	100	29	78	25	25	QANCHOR_REFONLY
9	ezrasmink	5	96	23	63	21	308	UAms_opus-mt-en-fr-finetuned
10	madhangi	1	100	36	52	20	337	PixelPair_groq_llama
11	ezrasmink	5	96	22	62	18	345	UAms_marian-ft_content0.0
12	ezrasmink	5	96	22	62	18	347	UAms_marian-ft_content0.25
13	robinflier	2	100	20	65	15	352	UAms_MarianMT_CamemBERT
14	robinflier	2	100	19	63	14	356	UAms_MarianMT_XLMRoBERTa
15	n0shi	5	7	6	6	6	371	Satir_Classifier

3.2. Results

3.3. Official results

Table 2 presents the top-5 results per team for Task 2 using the location metric. According to the location metric, the team *dsgt-arc* [16] showed the best results. Note that this team was the best team in 2025 in the JOKER Task 2 Pun translation task [11, 12]. Team *n0shi* [19] showed the second-best result. Almost all runs from the team *arampageos* [20] have very similar scores and follow the best run of *n0shi* [19]. The second-best run of *n0shi* is closely followed by the runs of *ezrasmink* [17]. All submitted runs are much better than the identity baseline (copying source English text into target).

Table 3 shows the top-3 results per team according to the percentage of successful translations, i.e., translations containing wordplay and preserving the meaning of the source pun. The results closely match the ranking based on the location metric. The best team, according to the manual evaluation, is *dsgt-arc* [16], as in the case of the location metric. There are slight variations in the top-3 results for this team, but the scores are very close in both the location metric and manual evaluation. Note that only 96 translations were evaluated for these runs instead of 100, likely because the matching was performed using the English source text, and a few source strings might slightly vary in the runs. In total, 44 translations over the 96 evaluated ones were considered to be successful. Both runs *ensemble_0_0_0_100_all_pooled_transl* and *ensemble_0_0_0_100_gpt_pooled_transl* have the same number of successful translations but with a slight difference in the appreciation of wordplay (higher for *ensemble_0_0_0_100_all_pooled_transl*) and meaning preservation (higher for *ensemble_0_0_0_100_gpt_pooled_transl*). The runs of *dsgt-arc* [16] are by far the runs with the largest proportion of wordplay. However, the runs of *dsgt-arc* [16] achieve the lowest meaning preservation scores. Their translations are often highly creative but may alter the original joke's meaning. For example, the pun *Camille relocated to Little Italy: They made her a cougher* is translated as *Quand Camille a rejoint la mafia de Little Italy, ils ont fait d'elle une vraie tousseuse à gages* ("When Camille joined the Little Italy Mafia, they made her a real hitwoman"), where "tousseuse à gages" is a creative play on "tueuse à gages" ("contract killer"). While the wordplay is successfully recreated, the original meaning is not fully preserved. The top-3 runs of *dsgt-arc* [16] outperform the second-best team, *n0shi* [19], by more than a factor of two in terms of the proportion of translations containing wordplay.

The second-best team is *n0shi* [19] with the run *Satir_CVO_Direct_mistral-large:123b* as in the case of the location metric. Their second-best run *Satir_Baseline* outperforms the best runs of the team

Table 4
Top-5 results per team for Task 2 Pun Translation (BLEU). BLEU refers to the BLEU Score, P1, P2, P3, and P4 are BLEU Precision 1, 2, 3, and 4 respectively. # runs is the number of runs submitted by each team.

# team	# runs	BLEU	P1	P2	P3	P4	Rank	run_id
1 arampageos	340	40.61	64.21	44.49	34.65	27.48	1	FINAL_REBUILD_319277
2 arampageos	340	40.61	64.21	44.49	34.65	27.48	2	FINAL_DROP_105_106
3 arampageos	340	40.60	64.19	44.48	34.64	27.46	3	FINAL_DROP_51_60_71_80
4 arampageos	340	40.60	64.19	44.47	34.64	27.46	4	FINAL_DROP_71_80
5 arampageos	340	40.60	64.19	44.48	34.64	27.46	5	FAST_BEST_PLUS_105_109
6 n0shi	5	38.29	63.37	42.72	32.19	24.68	320	Satir_CVO_Direct_mistral-large:123b
7 ezrasmink	5	35.77	60.95	39.88	29.69	22.67	333	UAms_marian-ft_content0.25
8 ezrasmink	5	35.76	60.97	39.88	29.67	22.66	334	UAms_marian-ft_content0.0
9 ezrasmink	5	35.68	60.59	40.01	29.66	22.54	335	UAms_nllb-ft_content0.25
10 robinflier	2	35.13	60.49	39.44	29.08	21.96	336	UAms_MarianMT_CamemBERT
11 robinflier	2	35.03	60.65	39.33	28.91	21.84	337	UAms_MarianMT_XLMRoBERTa
12 ezrasmink	5	34.49	60.43	38.74	28.34	21.32	340	UAms_opus-mt-en-fr-finetuned
13 madhangi	1	30.97	55.81	34.94	25.27	18.67	343	PixelPair_groq_llama
14 n0shi	5	27.78	54.31	31.81	22.02	15.66	347	Satir_Baseline
15 ezrasmink	5	27.31	54.72	31.68	21.40	14.99	348	UAms_mbart-ft_content0.25
16 n0shi	5	24.81	51.89	28.77	19.16	13.24	349	Satir_Classifier
17 dsgt-arc	37	21.54	47.32	25.17	16.35	11.05	350	ensemble_0_0_0_100_gemini_pooled
18 dsgt-arc	37	21.53	46.85	25.00	16.37	11.20	351	gemini_3.1_pro_preview_single
19 n0shi	5	21.52	46.74	24.63	16.35	11.39	352	Satir_BaselineCVO
20 dsgt-arc	37	20.72	46.38	24.41	15.61	10.43	353	ensemble_0_0_0_100_all_pooled
21 dsgt-arc	37	19.57	45.08	23.08	14.57	9.68	354	ensemble_0_0_0_100_gpt_pooled
22 dsgt-arc	37	18.85	43.85	22.05	13.95	9.37	355	ensemble_15_15_35_35_all_pooled
23 n0shi	5	17.92	42.37	20.82	13.23	8.83	360	Satir_ClassifierCVO

arampageos [20] which differs slightly from the ranking based on the location metric. Similarly, only 85 translations were evaluated for this run, which may have led to a slight underestimation of its performance. The runs of *arampageos* [20] are the most faithful to the English source texts, with by far the highest proportion of translations preserving the original meaning. Almost all runs of *arampageos* [20] are scored very similarly, and they are followed by the *UAms_opus-mt-en-fr-finetuned* run of *ezrasmink* [17].

Table 4 shows the results on the test set in terms of BLEU. Although the runs of *arampageos* [20] achieve the highest automatic scores, they do not rank among the top-performing runs according to either the manual evaluation or the location-based metric. Overall, this ranking is almost the reverse of those produced by the manual evaluation and the location-based metric, with the runs of *dsgt-arc* [16] appearing near the bottom of the table.

The location-based metric shows a strong agreement with manual evaluation (number of successful translations), achieving a Spearman rank correlation of 0.8517 for the rankings of the top-5 runs per team. These results are consistent with the high correlation between these two evaluation approaches last year [12]. In contrast, the Spearman rank correlation between BLEU scores and the number of successful translations is -0.0841, demonstrating that BLEU is not predictive of successful wordplay translation.

3.4. Train Results

Following previous editions of the shared task, participants submitted runs for both the training and test sets, allowing us to analyze potential overfitting and other dataset-specific effects. The participants' results on the training set, evaluated using the location-based metric and BLEU, are presented in Table 5 and Table 6 respectively.

The best-score runs of *arampageos* [20] are merely 100% in terms of BLEU Score and all BLEU

Table 5
Top-5 results per team for Task 2 Pun Translation (pun location metric) on the **Train data**

#	team	# runs	Score	Rank	run_id
1	arampageos	340	87.05	1	TrainQrelsLeak_EXACT_ALL_MEDOID
2	arampageos	340	86.98	2	TrainQrelsLeak_FUZZY95_LOCKED_MEDOID
3	arampageos	340	86.98	3	TrainQrelsLeak_FUZZY94_LOCKED_MEDOID
4	arampageos	340	86.98	4	GemmaChunk_21_40
5	arampageos	340	86.98	5	TrainQrelsLeak_FUZZY92_LOWER_LOCKED_MEDOID
6	ezrasmink	5	46.33	94	UAms_nllb-ft_content0.25
7	robinflier	2	45.27	95	UAms_MarianMT_CamemBERT
8	ezrasmink	5	44.91	96	UAms_opus-mt-en-fr-finetuned
9	ezrasmink	5	44.91	97	UAms_marian-ft_content0.0
10	ezrasmink	5	44.91	98	UAms_marian-ft_content0.25
11	n0shi	5	44.27	99	Satir_Baseline
12	robinflier	2	43.13	100	UAms_MarianMT_XLMRoBERTa
13	n0shi	5	42.99	101	Satir_BaselineCVO
14	dsgt-arc	37	42.42	102	ensemble_0_0_0_100_all_pooled_transl
15	dsgt-arc	37	41.64	103	ensemble_0_0_0_100_gpt_pooled_transl
16	dsgt-arc	37	41.64	104	gpt_5.5_single
17	dsgt-arc	37	41.57	105	ensemble_0_0_0_100_gemini_pooled_transl
18	dsgt-arc	37	40.36	107	gemini_3.1_pro_preview_single
19	n0shi	5	39.00	117	team1_ClassifierCVO
20	ezrasmink	5	39.00	119	UAms_mbart-ft_content0.25
21	n0shi	5	38.24	123	team1_TranslateByDecomposition
22	n0shi	5	37.37	323	Satir_CVO_Direct_mistral-large:123b
23	madhangi	1	28.75	381	PixelPair_groq_llama

precision. Their scores on the location metric are around 86-87%, suggesting strong overfitting, since these runs do not appear among the best on the test set.

The runs of *ezrasmink* [17] follow the runs of *arampageos* [20] with BLEU Score over 50% and location-based score around 45%. Other runs of the University of Amsterdam submitted by *robinflier* [17] have very close scores. On the test set, BLEU scores decrease to approximately 35%, while the location-based scores drop by up to 25%. This performance degradation may be explained by the translation models being fine-tuned on the training data, although a slight degree of overfitting of the pun detector may also have contributed.

The runs of *n0shi* [19] show less overfitting, although their performance is better on the training data in both BLEU and the location metric.

The submissions of *dsgt-arc* [16] achieve almost identical BLEU scores on the training and test sets. In contrast, their location-based score decreases from 42% on the training set to 32% on the test set. One possible explanation is the smaller number of reference translations available for the test set. On average, each English pun in the training set is associated with four reference translations, whereas the test set contains only 2.6 reference translations per English pun. Their highly creative translations may result in pun locations different from our references. Although these translations may still be judged successful by humans, the location-based metric rewards agreement with the reference locations and may therefore underestimate the performance of systems producing more diverse or original solutions. This effect is likely to be stronger on the test set, which contains only 2.6 reference translations per English pun on average, compared with four in the training set.

3.5. Manual Analysis

Table 7 provides the distribution of wordplay and meaning preservation in the 948 manually annotated French translations of 100 English puns. Overall, 43.35% of the translations were judged to contain wordplay. Only 27.43% were considered fully successful, i.e., they both preserved the meaning of the

Table 6
Top-5 results per team for Task 2 Pun Translation (BLEU) on the **Train data**

# team	# runs	BLEU	P1	P2	P3	P4	Rank	run_id
1 arampageos	340	99.72	99.87	99.77	99.68	99.59	1	TrainQrelsLeak_EXACT_ALL_MEDOID
2 arampageos	340	99.55	99.74	99.60	99.49	99.38	2	TrainQrelsLeak_FUZZY96_LOCKED_M
3 arampageos	340	99.55	99.74	99.60	99.49	99.38	3	TrainQrelsMulti_WRATIO96_ME
4 arampageos	340	99.55	99.74	99.60	99.49	99.38	4	TrainQrelsLeak_EXACT_LOCKED_ME
5 arampageos	340	99.55	99.74	99.60	99.49	99.38	5	TrainQrelsLeak_FUZZY94_LOCKED_M
6 ezrasmink	5	54.40	74.19	58.23	48.90	41.44	94	UAms_nllb-ft_content0.25
7 ezrasmink	5	53.45	74.06	57.04	47.70	40.50	95	UAms_opus-mt-en-fr-finetuned
8 ezrasmink	5	51.94	72.87	55.65	46.14	38.91	96	UAms_marian-ft_content0.0
9 ezrasmink	5	51.94	72.85	55.64	46.15	38.90	97	UAms_marian-ft_content0.25
10 robinflier	2	50.78	72.10	54.60	44.96	37.55	98	UAms_MarianMT_CamemBERT
11 robinflier	2	50.30	72.07	54.13	44.39	36.95	99	UAms_MarianMT_XLMRoBERTa
12 ezrasmink	5	48.35	70.31	52.14	42.44	35.12	103	UAms_mbart-ft_content0.25
13 n0shi	5	42.55	67.48	46.69	36.20	28.74	339	Satir_CVO_Direct_mistral-large:123b
14 n0shi	5	38.56	62.60	41.30	32.16	26.59	344	Satir_Baseline
15 madhangi	1	36.56	62.28	40.40	30.33	23.41	345	PixelPair_groq_llama
16 n0shi	5	34.80	56.70	36.71	28.97	24.31	346	Satir_BaselineCVO
17 n0shi	5	30.86	56.84	34.04	24.75	18.94	350	team1_TranslateByDecomposition
18 n0shi	5	30.42	54.25	32.91	24.56	19.52	351	team1_ClassifierCVO
19 dsgt-arc	37	25.45	52.95	29.35	19.71	13.70	352	gemini_3.1_pro_preview_single
20 dsgt-arc	37	23.98	52.02	27.96	18.35	12.39	353	ensemble_0_0_0_100_gemini_pooled
21 dsgt-arc	37	22.04	49.74	25.76	16.60	11.09	354	ensemble_0_0_0_100_all_pooled
22 dsgt-arc	37	20.20	47.83	23.80	14.89	9.83	355	ensemble_0_0_0_100_gpt_pooled
23 dsgt-arc	37	20.05	47.46	23.50	14.80	9.79	356	ensemble_25_25_25_25_gemini

Table 7
Distribution of wordplay and meaning preservation in the manually annotated data

Meaning preservation \ Wordplay	No	Yes	All
No	61 (6.43%)	33 (3.48%)	94 (9.92%)
Partially	151 (15.93%)	118 (12.45%)	269 (28.38%)
Yes	325 (34.28%)	260 (27.43%)	585 (61.71%)
All	537 (56.65%)	411 (43.35%)	948 (100.00%)

original pun and contained wordplay. This proportion increases to 40% when translations with partial meaning preservation are also considered successful. This represents a substantial improvement over the results of previous JOKER campaigns.

Table 8 reports the most frequent translation error categories observed in the manual analysis of 598 French translations of 93 English puns, together with representative examples. Note that some translations were annotated with 2 or more errors. The most common error was the loss of the original pun or cultural reference (316 instances), followed by literal translations that preserved the surface meaning but failed to recreate the wordplay (180 instances). Partial literal translations (59) and partially random translations (44) were also relatively frequent, indicating that systems often attempted to adapt the joke but achieved only partial success. Other error types, such as untranslated words, grammatical errors, missing words, and completely random translations, were comparatively rare.

Table 8
Most frequent translation error categories identified during the manual analysis, with representative examples.

Error type	#	EN	Erroneous FR	Reference FR
Lost pun/reference	316	Camille relocated to Little Italy: They made her a cougher	Camille a déménagé en Italie : ils l'ont faite tousseuse	Camille déménagea à Little Italy : ils lui firent une gaufre qu'elle ne pouvait pas refuser.
Literal translation	180	If you don't know how to choose music ask a guitarist – they know how to pick.	Si tu ne sais pas choisir la musique, demande à un guitariste, il sait choisir.	Pour choisir entre deux morceaux, prends un guitariste : c'est le meilleur médiator.
Partially literal translation	59	OLD DIVERS never die, they just get board.	Les vieux plongeurs ne meurent jamais, ils font seulement un trou dans l'eau.	Les vieux plongeurs ne meurent jamais : ils finissent juste par faire la planche.
Partially random	44	Holograms: the show must ghost on!	Hologrammes : le spectacle doit fantôme sur !	Hologrammes : le spectre-acle doit continuer !
Other	23	When walking through a pig barn be careful how you maneuver.	Quand vous traversez une porcherie, faites attention à comment vous manœuvrez pour éviter le fumier.	En traversant une porcherie, faites gaffe à comment vous manœuvrez… ça pourrait sentir le fumier.
Partially untranslated	18	Flying these days is a frisky business.	Voler aujourd'hui, c'est une affaire frisky.	Prendre l'avion de nos jours, ça donne vraiment des palpations.
Incorrect grammar/-spelling	10	Curiouser and curiouser!	De curio en curio !	De plus très-curieux en plus très-curieux !
Missing words/punctuation	1	Henry perfected the Ford down to a T.	Henri a amélioré la Ford jusqu'à la lettre.	Henry a perfectionné sa Ford de A à T.
Random	5	Blossom into a new you!	Quel est le comble pour un nouveau toi ?	Épanouis-toi : il est temps de fleurir sans te planter.

3.6. Case Study

Generating an effective pun in French remains difficult for current LLMs. For example, in *The diners were fully sated, unaware that the wurst was yet to come,* the humor relies on the phonetic similarity between *wurst* and *worst*. The model translates it as *Les dîneurs étaient repus, ne sachant pas que le pire-saucisse allait arriver*. Rather than recreating the wordplay, it literally translates *worst (pire)* and *wurst (saucisse)* and merges them into the artificial compound *pire-saucisse*. While this indicates that the model recognizes the presence of wordplay, the resulting expression is neither natural nor humorous in French.

A similar phenomenon can be observed with *What do you call a tissue that is sleeping? A napkin,* where the joke depends on the relationship between nap and napkin. The translation, *Comment appelle-t-on un mouchoir qui dort ? Un mouchoir dormant*, preserves only the literal meaning of the joke. Again, the lexical relationship between *nap* and *napkin*, which creates humor, is lost in the translation. As a result, the French version preserves only the literal meaning, with the wordplay disappearing entirely.

Some puns can be translated successfully through a straightforward literal translation. For example, *In order to lose weight, the publisher has decided to stop devouring his volumes* is translated as *Pour perdre du poids, l'éditeur a décidé d'arrêter de dévorer ses volumes*. Here, the ambiguity of *devouring volumes* is preserved because both English and French allow the verb *devour* to refer either to reading books eagerly or to eating excessively. Consequently, the translation retains both the original meaning and the wordplay. Notably, all translations of this joke in our manually evaluated examples successfully preserved both aspects through this literal translation strategy.

Another example of literal translation is the one of *One of the reasons more politicians should be drinking this hot beverage is because it will give them honest-TEA*, which is translated as *Les politiciens devraient boire plus de thé car ça leur confèrerait plus d'honnête-THÉ*. Because *tea* and *thé* are direct cognates, the model can preserve both the meaning and the wordplay through a nearly literal translation.

When no such correspondence exists, however, literal translation is no longer sufficient. For instance, *He sold candy and chocolate. A lot of girls were sweet on him* is translated as *Il vendait des bonbons et du chocolat ; pas étonnant que toutes les filles le trouvent à croquer*. Here, the original pun on *sweet* (meaning both *sugary* and *attracted to*) cannot be reproduced directly in French. Instead, the model replaces it

with the idiomatic expression *à croquer* (literally, *good enough to eat*), which preserves the humorous intent and semantic connection to sweets while recreating the wordplay in a form that is natural in French.

Sometimes, the models succeed in producing a pun only by altering the meaning of the original joke. Although the translation remains humorous, it deviates from the source text. A striking example is *In medical school he worried about passing as a surgeon, but he made the cut*, which is translated as *Au concours de coiffure, il craignait de ne pas réussir sa coupe, mais il a remporté la coupe.* Rather than preserving the original setting of medical school and surgery, the model shifts the scene to a hairdressing competition, where the ambiguity of *coupe* (*haircut* and *trophy/cup*) creates a new wordplay. While the translation is humorous and contains a successful wordplay, it no longer conveys the same situation or meaning as the source joke.

4. Discussion and Conclusions

In this paper, we presented an overview of Task 2 *Translate Puns from English to French.* Previously, we constructed the training data which includes 1,405 instances of wordplay in English and 5,838 professional human translations into French [7]. This year, we used new 6,951 reference translations of 2,656 English puns as test data. Since its launch in 2022, the task has led to the creation of one of the largest available parallel corpora for pun translation, which has nearly tripled in size and now includes annotations for wordplay locations. This continuously expanding resource provides a valuable benchmark for evaluating and advancing automatic pun translation.

We evaluated runs using a new location-based pun metric. The location-based metric closely matches the manual evaluation, achieving a strong Spearman correlation of 0.8517, while BLEU shows virtually no correlation with translation success ($\rho = -0.0841$), confirming that lexical overlap is not an appropriate measure for this task.

Since the introduction of the task, we have observed considerable improvements in the quality of automatic pun translation. The 2025 edition marked a significant shift from predominantly literal translations toward approaches that actively recreate wordplay. In particular, systems trained on the JOKER corpus demonstrated that non-literal, creativity-oriented translation strategies can substantially improve the quality of pun translations, with the best approach producing valid translations in approximately 77% of cases [13, 12]. This year, participating systems were increasingly able to recreate wordplay rather than relying solely on literal translations.

14 teams submitted 446 runs to CodaBench. Four teams submitted working notes for this task.

The *dsgt-arc* [16] submissions consistently achieved the best performance by generating substantially more successful wordplays, although this creativity sometimes came at the expense of meaning preservation. In contrast, the *arampageos* [20] systems produced the most faithful translations of the source meaning but recreated wordplay much less frequently, highlighting the trade-off between semantic fidelity and humorous creativity.

A comparison of the training and test results revealed varying degrees of overfitting across the participating systems. While the retrieval-based runs of *arampageos* [20] achieved near-perfect BLEU scores on the training set, their performance dropped substantially on the test set, suggesting strong dataset-specific optimization. In contrast, the *dsgt-arc* [16] systems generalized more consistently, although their highly creative translations may have been underestimated by the location-based metric on the test set due to the smaller number of reference translations and the resulting lower coverage of pun locations.

The manual evaluation shows a clear improvement over previous editions of the JOKER task. Overall, 43.35% of the translations contained wordplay, with 27.43% fully preserving both the original meaning and the wordplay; this rises to 40% when partially preserved meanings are also considered successful. The most frequent errors were the loss of the original pun or cultural reference, followed by literal translations that preserved the meaning but failed to recreate the humor. Other error types, such as untranslated words, grammatical mistakes, or random translations, were comparatively rare. Literal

translation is effective only when the source and target languages share similar lexical or semantic ambiguities, whereas language-specific puns typically require creative reformulation. Some models can generate natural French wordplay by replacing the original joke with an equivalent one, but this creativity may come at the cost of preserving the original meaning.

For more information about the JOKER lab this year and other tasks, please refer to the overview paper [1], and the Working Notes papers for Task 1: Humour-aware Information Retrieval [2], Task 3: Onomastic Wordplay Translation [3], and Task 4: Humour Generation [4]. Visit the JOKER website at https://joker-project.com for any other information related to the track.

Acknowledgments

We thank the Master's students in translation and technical writing from the University of Brest for participating in data annotation, and CodaBench for hosting the competition.

This project is funded by the National Research Agency (ANR-19-GURE-0001) under the program "*Investissements d'avenir*" integrated into France 2030. Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO NWA #1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, et al., Overview of the CLEF 2026 JOKER track: Humor detection, search, and translation, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[2] P. Vachharajani, A. Mukherjee, J. Singh, K. Tewari, S. Pal, J. Kamps, L. Ermakova, Overview of the CLEF 2026 JOKER Task 1: Humor-Aware Information Retrieval in English and Hinglish, in: [21], 2026.

[3] L. Ermakova, Y. Naud, L. Binet, J. Kamps, Overview of the CLEF 2026 JOKER Task 3: Onomastic Wordplay Translation from English to French, in: [21], 2026.

[4] I. Kuzmin, A.-G. Bosser, J. Kamps, L. Ermakova, Overview of the CLEF 2026 JOKER Task 4: Humor Generation, in: [21], 2026.

[5] L. Ermakova, T. Miller, F. Regattin, A. Bosser, C. Borg, É. Mathurin, G. L. Corre, S. Araújo, R. Hannachi, J. Boccou, A. Digue, A. Damoy, B. Jeanjean, Overview of joker@clef 2022: Automatic wordplay and humour translation workshop, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 447–469. URL: https://doi.org/10.1007/978-3-031-13643-6_27. doi:10.1007/978-3-031-13643-6_27.

[6] L. Ermakova, F. Regattin, T. Miller, A. Bosser, C. Borg, B. Jeanjean, É. Mathurin, G. L. Corre, R. Hannachi, S. Araújo, J. Boccou, A. Digue, A. Damoy, Overview of the CLEF 2022 JOKER task 3: Pun translation from english into french, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast

(Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1681–1700. URL: https://ceur-ws.org/Vol-3180/paper-128.pdf.

[7] L. Ermakova, A. Bosser, A. Jatowt, T. Miller, The JOKER corpus: English-french parallel data for multilingual wordplay recognition, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2796–2806. URL: https://doi.org/10.1145/3539618.3591885. doi:10.1145/3539618.3591885.

[8] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER - CLEF-2023 track on automatic wordplay analysis, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 397–415. URL: https://doi.org/10.1007/978-3-031-42448-9_26. doi:10.1007/978-3-031-42448-9_26.

[9] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 2 - pun location and interpretation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1804–1817. URL: https://ceur-ws.org/Vol-3497/paper-150.pdf.

[10] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 3 - pun translation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1818–1827. URL: https://ceur-ws.org/Vol-3497/paper-151.pdf.

[11] L. Ermakova, R. Campos, A. Bosser, T. Miller, Overview of the CLEF 2025 JOKER lab: Humour in machine, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 315–337. URL: https://doi.org/10.1007/978-3-032-04354-2_18. doi:10.1007/978-3-032-04354-2_18.

[12] L. Ermakova, R. Campos, A. Bosser, T. Miller, Overview of the CLEF 2025 JOKER task 2: Wordplay translation from english into french, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2761–2779. URL: https://ceur-ws.org/Vol-4038/paper_219.pdf.

[13] R. Taylor, B. Herbert, M. Sana, Pun intended: Multi-agent translation of wordplay with contrastive learning and phonetic-semantic embeddings for CLEF JOKER 2025 task 2, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 2870–2888. URL: https://ceur-ws.org/Vol-4038/paper_229.pdf.

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

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

[16] R. Taylor, A. Brikman, P. Awate, Searching for Sound-Meaning Collisions: Graph-Based Affordance

Retrieval and Multi-Evaluator Ranking for Cross-Lingual Pun Translation at CLEF 2026 JOKER Task 2, in: [21], 2026.

[17] J. van Bakel, R. Flier, E. Smink, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2026 JOKER Track, in: [21], 2026.

[18] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 1 - pun detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1785–1803. URL: https://ceur-ws.org/Vol-3497/paper-149.pdf.

[19] R. Pylypiv, M. Kuznietsov, M. Szabóová, TUKE Satira at the CLEF 2026 JOKER Track: Constant-Variable Optimization for English-French Pun Translation, in: [21], 2026.

[20] G. Arampatzis, A. Arampatzis, DUTH at CLEF JOKER 2026: Conservative Hybridisation for Humor-Aware Retrieval, Pun Translation, Onomastic Wordplay Translation, and Multilingual Humour Generation, in: [21], 2026.

[21] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.