Overview of the CLEF 2025 JOKER Task 2: Wordplay Translation from English into French

Liana Ermakova[1], Anne-Gwenn Bosser[2], Tristan Miller[3,4,*] and Ricardo Campos[5,6]

[1]*Université de Bretagne Occidentale, HCTI, France*
[2]*Bretagne INP - ENIB, Lab-STICC CNRS UMR 6285, France*
[3]*Department of Computer Science, University of Manitoba, Winnipeg, Canada*
[4]*Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria*
[5]*INESC TEC, Porto, Portugal*
[6]*University of Beira Interior, Covilhã, Portugal*

Abstract

This paper describes Task 2 of the CLEF 2025 JOKER track on the translation of puns from English into French. We outline the overall structure and setup of the shared task, discuss the approaches employed by the participants, and present and analyse the results they achieved. We also describe experiments with a promising new approach for the automatic evaluation of pun translation. Despite the significant improvements observed this year by participating systems, most of which used state-of-the-art large language models, we find wordplay translation to remain a complex and demanding task. Among the manually evaluated translations, 37.5% successfully preserved the meaning and involved wordplay, with success rates per English pun varying widely.

Keywords

wordplay, puns, computational humour, machine translation

1. Introduction

This paper presents an overview of Task 2 of the CLEF 2025 JOKER[1] evaluation campaign [1, 2], where participants are tasked with automatically translating puns across languages. The overall objective of the JOKER track series [1, 3, 4, 5], which started in 2022, is to facilitate collaboration among linguists, translators, and computer scientists to advance the development of automatic interpretation, generation, and translation of wordplay. The pun translation task has been a staple of JOKER since its inception; this year it was joined by tasks on humour-aware information retrieval [6] and onomastic wordplay translation [7].

A pun is a form of wordplay that exploits multiple meanings of a word, or words with similar sounds but different meanings. Puns pose challenges in translation, as they often rely on language-specific nuances that may not have direct equivalents in other languages. Nonetheless, it can be important to preserve wordplay in the target text, even if the exact type of wordplay or the specific meaning is changed. In Task 2, the goal is to translate English punning jokes into French in a way that preserves, as much as possible, both the form and meaning of the original. For example, "I used to be a banker but I lost interest" might be rendered into French as "*J'ai été banquier mais j'en ai perdu tout l'intérêt*". This fairly straightforward translation preserves the pun, since *interest* and *intérêt* happen to share the same double meaning.

Previous iterations of this task in JOKER have seen extremely low success rates, even for systems making use of state-of-the-art large language models (LLMs). For example, in JOKER-2023's manual evaluations, the highest success rate of English–French translations preserving both the form and sense

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
0000-0002-7598-7474 (L. Ermakova); 0000-0002-0442-2660 (A. Bosser); 0000-0002-0749-1100 (T. Miller); 0000-0002-8767-8126 (R. Campos)

[1] https://www.joker-project.com

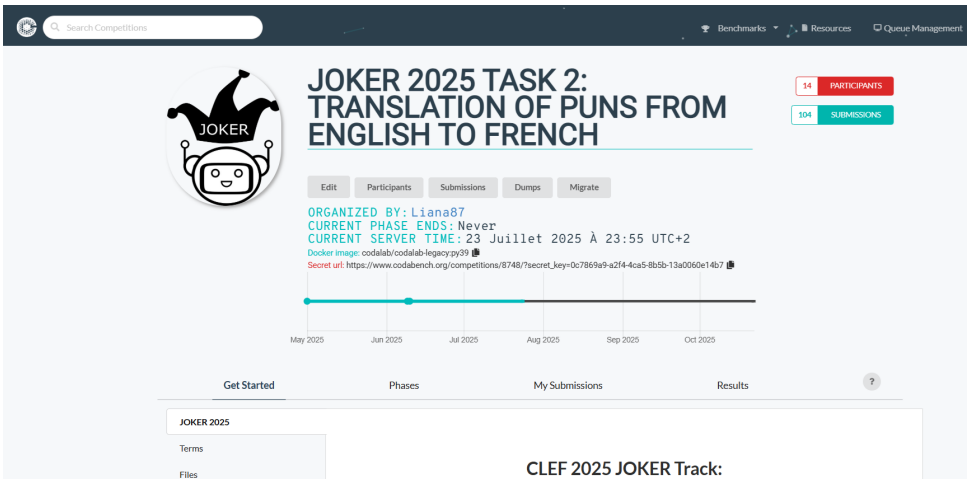


Figure 1: CLEF 2025 JOKER Task 2 on Codabench

Table 1
Number of runs submitted for Task 2

team	#
alecs [8]	8
arampageos [9]	12
cryptix [10]	1
igoranchik [11]	13
kamps [8]	2
pjmathematician [12]	4
rdtaylorjr [13]	4
sarath_kumar [10]	5
verbanex [14]	3
Total	52

of the original wordplay was only 6%. This highlights the need for increased community focus on this task.

This year nine teams submitted 52 runs for Task 2, reflecting the community's stable interest automatic pun translation. Table 1 shows the number of runs submitted by each participating team.

One of the innovations made in CLEF 2025 JOKER was infrastructural: for the first time, we ran the task on Codabench[2] [15], the Free Software web-based platform for organising AI benchmarks (see Figure 1). Codabench greatly facilitated running the track in 2025 and attracted many new participants, all of whom had full access to the competition, including the submission and leaderboard pages. We continue to receive new registrations and post-competition submissions. However, in this paper we present only runs submitted before the official results were communicated to the participants.

In the remainder of this paper, we describe the data used in Task 2 (Section 2), the evaluation measures (Section 3), and the participants' approaches (Section 4), and then present an analysis of their results for both the training and test data (Section 5). In addition to traditional machine translation evaluation measures, such as BLEU [16] and BERTScore [17], we examined the participants' performance using the dataset we created to identify words or phrases that have multiple meanings (pun locations) for the CLEF 2023 JOKER Task 2 [18, 4, 19]. We show that this approach is promising to evaluate translation of wordplay based on multiple meanings. Section 6 concludes the paper.

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

Table 2
Histogram of the number of references and locations per English pun in the Task 2 training data

#	references	locations
1	396	578
2	237	254
3	234	192
4	169	110
5	88	70
6	55	58
7	42	31
8	27	15
9	25	15
10	12	11
11	13	20
12	20	12
13	7	7
14	8	6
15	12	8
16	10	7
17	6	2
18	8	3
19	7	4
20	6	2
>20	11	—

2. Data

2.1. Training data

The training data for Task 2, which builds on previous editions [3, 20, 4, 5], consists of 1,405 instances of wordplay in English, with a total of 5,838 French translations sourced from human professionals with manually annotated words or phrases with multiple meanings (pun locations). Pun location annotations were collected for previous JOKER evaluation campaigns [19, 4, 18]. Table 2 shows a histogram of the number of references and distinct locations per English pun in the training data.

We provide training data in the format of JSON qrels files with the following fields:

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

Example of a training file:

```
[
    {
        "id_en":"en_1",
        "en":"I used to be a banker but I lost interest",
        "fr":"J'ai été banquier mais j'en ai perdu tout l'intérêt."
    }
]
```

Table 3
Histogram of the number of references and locations per English pun in the Task 2 test data

#	references	locations
1	1,252	1,382
2	172	220
3	133	68
4	53	10
5	39	1
6	22	1
7	8	—
8	2	—
9	1	—

2.2. Test data

For the 2025 edition, we collected 2,615 new manual translations of 1,682 distinct puns in English with manually annotated pun locations that we used for the test data. Some of the pun location annotations were collected for previous JOKER evaluation campaigns [19, 4, 18]. We expanded this data with annotations of new references. Table 3 shows a histogram of the number of references and distinct locations per English pun in the test data. For 25% of English puns from the test data, we have multiple references and multiple locations. There are 1,382 English puns with a single location, while only 1,252 of them have a single reference. However, we have much more multiple references and distinct locations on training data.

The test input data is provided in JSON format with the following fields:

- **id_en**: a unique identifier
- **en**: the text of the instance of source wordplay in English

An input example is as follows:

```
[
    {
        "id_en":"en_1",
        "en":"I used to be a banker but I lost interest"
    }
]
```

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_3_BLOOM
- **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":"team1_task_3_DeepL",
        "manual":0,
        "id_en":"en_1",
        "en":"I used to be a banker but I lost interest"
        "fr":"J'ai été banquier mais j'en ai perdu tout l'intérêt"
    }
]
```

3. Evaluation

We evaluated the runs with the following metrics:

BLEU (BiLingual Evaluation Understudy) computes the translation's overlap in vocabulary overlap with a reference translation [16]. We used the sacreBLEU implementation [21] with the default tokeniser 13a. We report the BLEU score (harmonic mean) and the BLEU precisions for n-grams for $n = 1, 2, 3, 4$.

BERTScore computes tokenwise similarity scores between the candidate translation and a reference translation using contextual embeddings [17]. We used the Python implementation from the bert-score package.[3] We report mean values of BERTScore precision, recall, and F_1 over all references.

Pun location–based evaluation allows for a more fine-grained analysis of generated translations. We checked for words or phrases with multiple meanings (pun locations) from the reference texts by combining French reference translations with pun location annotations from the dataset used for JOKER 2023's Task 2: Pun Location and Interpretation [4, 18, 19]. We completed this data with pun location annotations of the new references.

Manual evaluation consisted of human assessments of 1,297 French translations of 50 distinct source English puns in terms of meaning preservation and the presence of wordplay. This manual evaluation was performed by a Master's student in translation who specialises in wordplay translation and is a native French speaker.

4. Participants' approaches

Nine teams submitted 52 official runs for this task. Statistics on the runs are presented in Table 1. Team names are reported according to the participant names listed on Codabench. The approaches used were as follows:

arampageos [9] This team combined neural machine translation systems with a handcrafted translation dictionary of particularly challenging puns. The machine translation systems included Google Translate, Argos Translate, the Helsinki-NLP/opus-mt-en-fr models, Facebook's M2M100 (418M and 1.2B), MBART50, and NLLB (1.3B and distilled 600M). Their two-stage pipelines first checked whether the input pun matched the curated set, otherwise forwarding the input to the machine translation system.

verbanex [14] This team relied on extensive data preprocessing, including sentiment classification and phoneme conversion, to help the trained translation model capture emotional tone and pronunciation ambiguities. They used two different fine-tuning strategies – full parameter optimisation and parameter-efficient adaptation techniques – with the mBART-50 English-to-French translation model.

rdtaylorjr [13] This participant relied on a three-stage approach. The first stage consisted of training multiple LLMs (provided by openAI, Google, Mistral, or DeepSeek) using a contrastive learning approach. In addition to the training set we provided, they used data from the JOKER 2023 shared task on pun location and interpretation, as well as a contrastive learning dataset constructed by neutralising puns of their French dataset. The second stage of the approach is based on chain-of-thought prompting making use of semantic and phonetic embeddings for the French language. Finally, evaluator agents were used to iterate over various properties of the proposed translations (conserving literal/contextual meaning, emotion level, and understandability in the target language).

[3] https://pypi.org/project/bert-score/

alecs and kamps [8] These participants used a fine-tuned MarianMT sequence-to-sequence model, T5ForConditionalGeneration, T5-base, Meta AI NLLB-200-1.3B, and mBART-large-cc25.

pjmathematician [12] This team fine-tuned different Qwen models, including the Qwen2.5-14B, experimenting with different LoRA parameters on the provided corpus. They then used a simple prompting approach for requesting translations of puns.

igoranchik [11] This team used supervised fine-tuning with the aim of forcing a model to learn higher quality responses. They also used an Adaptive Rejection Preference Optimisation (ARPO) [22] implementation[4] in an attempt to enhance humour retention.

cryptix and sarath_kumar [10] These participants used back-translation for data augmentation. They fine-tuned the MarianMT model and used a loss function combining humour preservation metrics from a rule-based module evaluating humour preservation with standard BLEU metrics.

All participants who submitted runs also submitted system description papers to the Working Notes volume [23]. Two teams from the same university (alecs and kamps) submitted a single joint report, as did teams cryptix and sarath_kumar, resulting in a total of seven Working Notes from the participants of Task 2. Despite the requirement to include the team ID in the run name, participants' submissions often differed in their run names, registration details, and Codabench IDs. We manually matched the Working Notes with the submitted runs and report the results using the team names provided in those submissions.

5. Results

5.1. Test data

Tables 4, 5, and 6 report, respectively, the results based on BLEU, BERTScore, and the pun location–based metric for Task 2 on the test data. The tables show the run names as submitted, except that we remove the term *task_2* from the middle of the names to improve readability and avoid redundancy.

The top three runs according to BERTScore and BLEU – namely, UvA_finetunedNLLB-1.3B [8], Skommarkhos_Lucie_SFT [11], and Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4 [11] – are very close to each other. Interestingly, the fourth-best run according to BERTScore, UvA_finetuned-MarianMT [8], drops to 14th position according to BLEU, while the fourth-best run according to BLEU, Skommarkhos_Lucie_SFT_ARPO [11], drops to 12th position according to BERTScore. The top 14 results according to BLEU are shared by the teams Skommarkhos [11] and the University of Amsterdam [8]; for BERTScore these positions are also shared with the teams arampageos [9] and Cryptix [10]. The approaches of the teams Skommarkhos [11] and the University of Amsterdam [8] are based on fine-tuning.

According to the pun location–based metric, the best runs are dsgt_o4_mini_multi_agent_discriminator and dsgt_o4_mini_chain_of_thought_phonetic_embeddings [13], with respectively 156 and 132 translations with locations shared with references. It is followed by teamX_aug [14] and pjmathematician_Q25-14 [10], despite these runs placing in the second half of all runs according to BLEU and BERTScore. They are followed by the run UvA_finetunedMarianMT [8], which is also in the fourth place according to BERTScore.

The top five runs shared only 7–9% of locations with references. This number corresponds to the percentage of successful manually evaluated machine translations we reported previously [20, 4]. As we discussed previously [20, 24], the percentage of locations shared with references is similar to the percentage of successful wordplay translations.

[4] https://github.com/fe1ixxu/ALMA

Table 4
Task 2 results in terms of BLEU score and BLEU n-gram precision (test data)

run ID	Score	$n = 1$	$n = 2$	$n = 3$	$n = 4$
Skommarkhos_Lucie_SFT	43.33	65.05	46.98	37.59	30.67
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	43.20	64.73	46.74	37.50	30.69
UvA_finetunedNLLB-1.3B	42.55	64.74	46.26	36.70	29.83
Skommarkhos_Lucie_SFT_ARPO	42.48	63.76	45.92	36.86	30.17
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	42.26	64.44	46.13	36.47	29.42
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	42.15	63.33	45.50	36.56	29.95
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	42.14	63.38	45.53	36.54	29.90
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	42.14	63.37	45.55	36.56	29.87
Skommarkhos_Lucie-7B-Instruct-v1.1	42.14	63.43	45.54	36.51	29.88
Skommarkhos_Lucie-7B-Instruct-v1.1	42.12	63.41	45.54	36.50	29.86
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	42.11	63.33	45.46	36.51	29.91
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	42.00	63.29	45.41	36.40	29.74
UvA_finetunedNLLB-1.3B&finetunedroBERTa	41.80	63.86	45.49	36.01	29.17
UvA_finetunedMarianMT	41.19	63.37	44.74	35.31	28.76
duth_hybrid_fusion	41.11	63.45	44.62	35.17	28.70
yourteamid_marianmt_pun_postedit	41.01	63.40	44.52	35.07	28.58
duth_xanthi_helsinki	41.01	63.40	44.52	35.07	28.58
Cryptix	41.01	63.40	44.52	35.07	28.58
Cryptix_marianmt	40.98	63.36	44.49	35.04	28.55
duth_xanthi_GoogleTranslate_fallback	40.94	62.75	44.21	35.12	28.84
UvA_finetunedMarianMT&finetunedroBERTa	40.85	62.90	44.38	35.03	28.49
Cryptix	40.75	62.54	43.98	34.95	28.69
duth_google_flant5_fallback	40.74	62.60	43.99	34.92	28.65
duth_xanthi_GoogleTranslate	40.73	62.59	43.98	34.91	28.64
duth_xanthi_GoogleTranslate_fallback	40.73	62.60	43.99	34.91	28.63
duth_xanthi_argos	40.49	63.21	44.13	34.73	28.24
pjmathematician_Q25-14	39.08	62.57	43.10	33.19	26.05
pjmathematician_Q25-14	38.49	61.88	42.37	32.62	25.66
pjmathematician_Q25-14	38.24	61.56	42.12	32.40	25.46
UvA_finetunedT5-base	36.77	60.29	40.58	30.94	24.15
duth_xanthi_m2m100_1_2B	36.46	61.22	41.01	30.91	23.90
Skommarkhos_Croissant_SFT_ARPO	36.35	60.12	40.30	30.47	23.63
Skommarkhos_Croissant_SFT	36.20	59.93	40.22	30.35	23.47
UvA_T5-base&finetunedroBERTa	36.14	59.75	39.92	30.30	23.61
teamX_aug	33.86	54.03	37.18	28.70	22.80
duth_xanthi_mbart50	32.73	57.21	36.32	26.81	20.62
duth_xanthi_t5_base_gpu	32.44	56.04	36.26	26.82	20.33
duth_combined_m2m100	30.09	56.97	34.81	24.41	17.66
teamX_final	29.11	53.62	32.49	23.39	17.63
teamX_aug	28.63	55.38	33.30	22.66	16.07
dsgt_o4_mini_multi_agent_discriminator	21.41	46.61	25.26	16.35	10.92
UvA_mBARTcc25&finetunedroBERTa	18.20	40.86	21.09	13.74	9.26
duth_xanthi_bloomz3b_local	16.68	41.17	19.57	12.15	7.91
UvA_finetunedmBARTcc25	16.55	39.64	19.49	12.22	7.95
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	16.52	39.85	19.78	12.12	7.79
dsgt_simple_mistral_medium	14.94	37.47	17.74	10.90	6.88
dsgt_simple_o4_mini	8.15	29.13	9.80	5.17	2.99
Cryptix_finetunedmarian	0.37	13.75	0.82	0.13	0.02
Cryptix_rulebased	0.00	0.04	0.01	0.01	0.00
yourteam_rulebased	0.00	0.04	0.01	0.01	0.00

5.2. Manual evaluation

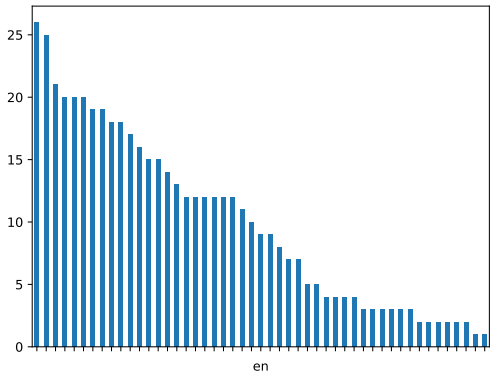
Among the translations evaluated manually (1,297 French translations of 50 distinct source English puns in total), we discovered 50 cases where the output was simply "[unk]", three empty translations, 72 untranslated texts, and four incomplete translations. Two outputs were a mix of English and French ("*Horloge en forme de the thinker qui annonce l'heure en disant I think it s 20.25 pm*" and "*I have to keep*

Table 5
Task 2 results in terms of BERTScore precision, recall, and F_1 (test data)

run ID	P	R	F_1
UvA_finetunedNLLB-1.3B	87.85	87.04	87.42
Skommarkhos_Lucie_SFT	87.74	87.15	87.42
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	87.61	87.01	87.28
UvA_finetunedMarianMT	87.72	86.82	87.24
UvA_finetunedNLLB-1.3B&finetunedroBERTa	87.55	86.96	87.23
UvA_finetunedMarianMT&finetunedroBERTa	87.50	86.78	87.11
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	87.31	86.91	87.08
duth_xanthi_GoogleTranslate_fallback	87.20	86.77	86.96
duth_xanthi_GoogleTranslate	87.20	86.77	86.96
duth_google_flant5_fallback	87.17	86.74	86.93
Cryptix	87.10	86.63	86.84
Skommarkhos_Lucie_SFT_ARPO	86.79	86.59	86.66
Skommarkhos_Lucie-7B-Instruct-v1.1	86.68	86.46	86.54
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	86.68	86.45	86.53
Skommarkhos_Lucie-7B-Instruct-v1.1	86.66	86.46	86.53
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	86.67	86.45	86.53
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	86.67	86.43	86.52
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	86.64	86.40	86.49
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	86.63	86.39	86.48
pjmathematician_Q25-14	87.00	85.95	86.45
UvA_finetunedT5-base	86.69	86.24	86.44
duth_hybrid_fusion	87.18	85.79	86.43
Cryptix_marianmt	87.17	85.77	86.42
duth_xanthi_argos	87.00	85.91	86.42
yourteamid_marianmt_pun_postedit	87.17	85.74	86.40
Cryptix	87.17	85.74	86.40
duth_xanthi_helsinki	87.17	85.74	86.40
pjmathematician_Q25-14	86.84	85.97	86.37
pjmathematician_Q25-14	86.71	85.89	86.27
UvA_T5-base&finetunedroBERTa	86.42	86.12	86.24
duth_xanthi_GoogleTranslate_fallback	86.34	85.91	86.10
Skommarkhos_Croissant_SFT	85.71	85.65	85.65
Skommarkhos_Croissant_SFT_ARPO	85.66	85.64	85.62
duth_xanthi_m2m100_1_2B	85.90	85.28	85.56
teamX_aug	85.69	85.46	85.46
duth_xanthi_mbart50	85.14	84.14	84.60
duth_combined_m2m100	84.76	84.05	84.37
teamX_aug	84.61	84.15	84.35
teamX_final	84.01	83.50	83.73
duth_xanthi_t5_base_gpu	83.91	83.60	83.71
dsgt_o4_mini_multi_agent_discriminator	79.84	81.57	80.66
UvA_mBARTcc25&finetunedroBERTa	80.07	80.00	80.00
UvA_finetunedmBARTcc25	79.48	79.79	79.59
duth_xanthi_bloomz3b_local	79.30	78.66	78.94
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	77.80	79.15	78.42
dsgt_simple_mistral_medium	77.60	79.12	78.30
Cryptix_finetunedmarian	75.06	73.56	74.27
dsgt_simple_o4_mini	73.82	73.95	73.85
yourteam_rulebased	64.26	54.35	58.86
Cryptix_rulebased	64.26	54.35	58.86

Table 6
Task 2 results in terms of the pun location–based metric (test data)

run ID	count	location	%
dsgt_o4_mini_multi_agent_discriminator	1682	156	9.27
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	1682	132	7.85
teamX_aug	1682	118	7.02
pjmathematician_Q25-14	1682	118	7.02
UvA_finetunedMarianMT	1682	114	6.78
UvA_finetunedMarianMT&finetunedroBERTa	1682	114	6.78
Cryptix	1682	113	6.72
Cryptix_marianmt	1682	113	6.72
duth_xanthi_helsinki	1682	113	6.72
yourteamid_marianmt_pun_postedit	1682	113	6.72
duth_xanthi_GoogleTranslate_fallback	1682	112	6.66
duth_hybrid_fusion	1682	112	6.66
Cryptix	1682	112	6.66
duth_google_flant5_fallback	1682	112	6.66
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	1682	111	6.60
pjmathematician_Q25-14	1682	111	6.60
duth_xanthi_GoogleTranslate	1682	111	6.60
duth_xanthi_GoogleTranslate_fallback	1682	111	6.60
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	1682	111	6.60
UvA_finetunedNLLB-1.3B&finetunedroBERTa	1682	111	6.60
duth_xanthi_argos	1682	109	6.48
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	1682	109	6.48
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	1682	109	6.48
Skommarkhos_Lucie-7B-Instruct-v1.1	1682	109	6.48
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	1682	108	6.42
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	1682	108	6.42
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	1682	108	6.42
Skommarkhos_Lucie-7B-Instruct-v1.1	1682	107	6.36
UvA_finetunedNLLB-1.3B	1682	107	6.36
UvA_T5-base&finetunedroBERTa	1682	107	6.36
pjmathematician_Q25-14	1682	107	6.36
UvA_finetunedT5-base	1682	106	6.30
Skommarkhos_Lucie_SFT_ARPO	1682	100	5.95
Skommarkhos_Croissant_SFT	1682	96	5.71
duth_xanthi_t5_base_gpu	1682	95	5.65
Skommarkhos_Croissant_SFT_ARPO	1682	92	5.47
duth_xanthi_m2m100_1_2B	1682	91	5.41
Skommarkhos_Lucie_SFT	1682	90	5.35
teamX_aug	1682	90	5.35
UvA_mBARTcc25&finetunedroBERTa	1682	89	5.29
teamX_final	1682	84	4.99
duth_xanthi_mbart50	1682	81	4.82
duth_combined_m2m100	1682	71	4.22
UvA_finetunedmBARTcc25	1682	64	3.80
dsgt_simple_mistral_medium	1682	60	3.57
duth_xanthi_bloomz3b_local	1682	45	2.68
Cryptix_finetunedmarian	1682	25	1.49
dsgt_simple_o4_mini	1682	18	1.07
yourteam_rulebased	1682	0	0.00
Cryptix_rulebased	1682	0	0.00



count	49
mean	9.94
std	7.12
min	1
25%	3
50%	9
75%	15
max	26

Figure 2: Number of successful translations per English pun

this fire alight, a crié Tom.") and seven had useless repetitions (e.g., "*Les vendeurs de chips ne peuvent pas vendre leurs produits, ils ne peuvent pas vendre leurs produits.*")

There were 572 translations that did not preserve the meaning of the source text, while 464 preserved it fully and 145 preserved it partially. There were 80 translations containing wordplay that did not preserve the meaning of the source pun (e.g., "Chips vendors don't get the dough unless their products sell." → "*Les agriculteurs ne font pas de blé s'il n'y a pas de blé dans les champs*", or "The scientist had trouble reducing the liquid, he just couldn't concentrate." → "*Le brasseur n'arrive pas à maintenir la mousse, pourtant il se fait mousser.*") Among the translations that preserved the meaning fully or partially, 487 involved wordplay, while 122 did not, resulting in 37.5% of successful translations over the total number of manually evaluated translations.

The number of successful translations per English pun is given in Figure 2. For half of source English puns the success is less than 25% with mean 10 and median 9 successful translation per source. The maximal number of successful runs per English source is 26 over 49. For 25% of English puns there is 3 or less successful translations, highlighting the inherent difficulty of rendering certain wordplays effectively across languages.

The only English pun without successful translation was "Having too many axe-like tools to do a particular job only adze to the confusion." Two English puns had only a single successful translation:

- "The geneticist taught his students how to mendel defective genes" → "*Le généticien a appris à ses étudiants à repriser leurs jeans… et leurs gènes !*" (dsgt_o4_mini_chain_of_thought_phonetic_-embeddings)
- "Volts – the dance you perform after an electric shock" → "*En anglais, le verbe « voltige » désigne une danse après une décharge électrique*" (duth_xanthi_bloomz3b_local)

The results of manual evaluation per run are given in Table 7. According to our manual evaluation, the best runs were dsgt_o4_mini_multi_agent_discriminator and dsgt_o4_mini_chain_of_thought_-phonetic_embeddings [13], achieving 37 and 36 successful translations, respectively. These results significantly outperformed the third-best run, which achieved only 26 successful translations. These results are consistent with the evaluation in terms of location-based metric. The runs with lowest performance according to the location-based metric are also low-scored by the expert.

The Pearson correlation coefficient between the manually attributed scores and the location-based metrics is 0.84, indicating a strong positive relationship between the two. The scatter plot in Figure 3 shows this relationship, including a regression line with a 95% confidence interval, which further illustrates the consistency of the association across the dataset. Such a high correlation suggests that the location-based evaluation captures key aspects of translation quality that align closely with expert judgments. Consequently, it can be considered a reliable proxy for assessing the quality of wordplay translation, offering a scalable and less resource-intensive alternative to manual evaluation. However,

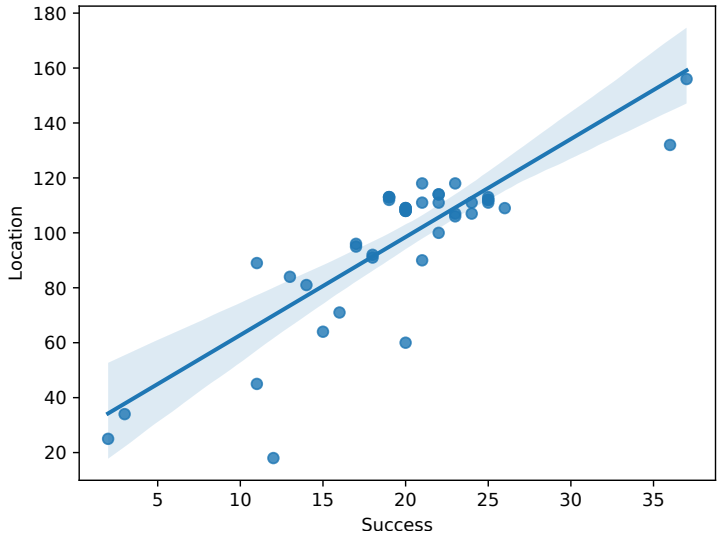


Figure 3: Scatter plot of manually attributed scores and location-based metrics with a regression line with a 95% confidence interval

further analysis is needed, as the current results are based on the official scores presented in Tables 6 and 7, and may not fully capture variability across different evaluators or contexts. Note that these results are drastically different from the ranking based on BLEU and BERT scores.

5.3. Training data

As in previous years, runs were submitted for both the training and test datasets in order to analyse potential overfitting and related effects. Tables 8, 9, and 10 report, respectively, the results based on BLEU, BERTScore, and the pun location–based metric for Task 2 on the training data.

Three runs (Cryptix_rulebased [10], pjmathematician_Q25-14 [12], yourteam_rulebased [10]) achieved a BLEU score of 100 on the training data, followed by 99.77 for teamX_final [14]. For the next four runs, all submitted by the University of Amsterdam (UvA_finetunedNLLB-1.3B, UvA_finetunedNLLB-1.3B&finetunedroBERTa, UvA_finetunedMarianMT, UvA_finetunedMarianMT&finetunedroBERTa) [8], we observe an almost twofold drop in performance. UvA_finetunedNLLB-1.3B remains high on the test data, ranking third according to the BLEU evaluation. Both rule-based runs (Cryptix_rulebased and yourteam_rulebased) are positioned at the bottom of the table in terms of test results. The best run in terms of BLEU on test set, Skommarkhos_Lucie_SFT [11] (43.33), has a BLEU score of 49.98 on the training data. pjmathematician_Q25-14 achieved BLEU scores of 100 on the training data and 39 on the test data. However, the same run ranked third-highest according to the location-based metric and fell within the second tier of manually ranked runs, highlighting a discrepancy between traditional BLEU evaluation and alternative assessment methods.

The BLEU score results on the training data have similar trend as the BERTScore (see Table 9). However, the top-scored teamX_final [14] on the training data has much lower rank on the test data according to BERTScore.

Cryptix_rulebased [10], pjmathematician_Q25-14 [12], and yourteam_rulebased [10] have 391 (27.83%) successful locations sharing the first rank on the training data closely followed by teamX_final [14] with the result 380 (27.05%). Cryptix_rulebased [10], yourteam_rulebased [10], and teamX_final [14] are in the bottom of the table according to the location-based metric on the test set suggesting overfitting on the training data. The next runs are dsgt_o4_mini_chain_of_thought_phonetic_embeddings and dsgt_o4_mini_multi_agent_discriminator [13] with results 184 (13.10%) and 183 (13.02%) respectively. These runs are the best according to manual evaluation and location-based metric on the test set, suggesting generalisation capacity. They are followed by the four runs of the University of Am-

sterdam [8] with slightly lower scores, and then teamX_aug [14]. Note that teamX_aug is ranked third according to the location-based metric on the test set. Thus, these two sets of results are comparable on the test and training data.

Note that it is problematic to directly compare the absolute scores between the training and test data, rather than run ranks, as the number of distinct locations per English pun is considerably higher in the training data than in the test data. (See Tables 3 and 2.)

6. Conclusion

In this paper, we have described the wordplay translation task of the JOKER track at CLEF 2025. This year, we expanded the corpus used in previous editions of the task [3, 19, 4] by introducing 1,682 new distinct source texts with 2,615 corresponding reference translations created by professional French native-speaker translators for the test purpose. We manually annotated pun locations in French translations in order to provide an automatic evaluation that takes into account pun ambiguity.

Nine teams submitted 52 runs for Task 2, demonstrating stable interest from the community in our perennial pun translation task. Participants used a variety of methods, including LLMs, commercial machine translation engines, out-of-the-box translation models, rule-based approaches, and various fine-tuning and training techniques to discriminate wordplay from non-wordplay.

We evaluated the participants' results using automated measures, specifically BLEU and BERT scores both on test and training sets. According to these automatic measures on the test data, the best results were achieved by the fine-tuned approaches of the teams Skommarkhos [11] and the University of Amsterdam [8]. Three runs – Cryptix_rulebased, pjmathematician_Q25-14, and yourteam_rulebased – achieved perfect BLEU scores of 100 on the training data, with another run scoring 99.77. However, both rule-based runs (Cryptix_rulebased and yourteam_rulebased) performed poorly on the test data, ranking near the bottom. The pjmathematician_Q25-14 run also showed signs of overfitting, achieving a BLEU score of 39 on the test data despite its perfect training score. The BLEU score results on the training data show a similar trend to those of BERTScore.

On the pun location–based metric, the best runs were dsgt_o4_mini_multi_agent_discriminator and dsgt_o4_mini_chain_of_thought_phonetic_embeddings [13]. These were followed by teamX_-aug [14] and pjmathematician_Q25-14 [10], which scored lower on BLEU and BERTScore, and then by UvA_finetunedMarianMT[8], which also ranked fourth by BERTScore. Rule-based runs achieved top scores on the training set but performed poorly on the test set, suggesting overfitting. In contrast, dsgt_o4_mini_multi_agent_discriminator and dsgt_o4_mini_chain_of_thought_phonetic_embeddings generalised well, ranking highest on both manual and location-based evaluations, followed by the University of Amsterdam runs and teamX_aug.

Manual evaluation confirmed dsgt_o4_mini_multi_agent_discriminator and dsgt_o4_mini_chain_-of_thought_phonetic_embeddings [13] as the best-performing runs, consistent with the location-based metric. Conversely, the lowest-ranked runs by this metric were also rated poorly by the expert, reinforcing the alignment between automated and manual evaluations.

Overall, we observe significant improvements in participants' results compared to previous years, based on both manual (up to 74% of successful translations for one team [13]) and location-based evaluations, particularly on the training data (up to 28% for rule-based appraoches on the training data). However, despite these significant improvements, wordplay translation remains a complex and demanding task. With the exception of runs exhibiting overfitting on the training data, the location-based evaluation results are consistent with those of previous years. Among the manually evaluated translations, 37.5% successfully preserved the meaning and involved wordplay, with success rates per English pun varying widely –half having under 25% good translations in French, a median of nine successful translations, and 25% with three or fewer – underscoring the difficulty of effectively rendering wordplay across languages.

One of the major obstacles in the development of wordplay machine translation is its evaluation. Destroying the wordplay may result in the text becoming nonsensical. The existing metrics do not take

into account punning words which can reward translations with completely lost sense. The strong Pearson correlation (0.84) between manual scores and the location-based metric indicates that the latter reliably reflects expert judgments. This suggests it can serve as a scalable, less resource-intensive proxy for evaluating wordplay translation quality. Manual scores and location-based metrics correlate closely but differ substantially from BLEU and BERTScore rankings, highlighting the limitations of the latter for evaluating wordplay translation. However, further analyses are needed. In future work, we will explore new perspectives on evaluating wordplay in machine translation based on the data constructed within the JOKER track.

Additional information on the track is available on the JOKER website: https://www.joker-project.com/

Acknowledgments

This work has received a government grant managed by the National Research Agency under the program Investissements d'avenir integrated into France 2030, with the Reference ANR-19-GURE-0001. It was also financed by National Funds through the Portuguese funding agency FCT through the project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020). Ricardo Campos would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC). We thank all other colleagues and students who participated in data construction, the translation contests, and the CLEF JOKER track.

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. Further, the authors used Gemini in order to: Generate images. 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, A.-G. Bosser, T. Miller, R. Campos, Overview of JOKER: Humour in the machine, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[2] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, CLEF 2025 JOKER lab: Humour in the machine, in: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2025, p. 389–397. URL: https://doi.org/10.1007/978-3-031-88720-8_59. doi:10.1007/978-3-031-88720-8_59.

[3] L. Ermakova, A.-G. Bosser, T. Miller, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of the CLEF 2024 JOKER track: Automatic humour analysis, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), volume 14959 of *Lecture Notes in Computer Science*, Springer, Cham, 2024, pp. 165–182. doi:10.1007/978-3-031-71908-0_8.

[4] L. Ermakova, T. Miller, A.-G. 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, volume 14163, Springer Nature Switzerland, Cham, 2023, pp. 397–415. doi:10.1007/978-3-031-42448-9_26.

[5] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, E. Mathurin, G. L. Corre, S. Araújo, J. Boccou, A. Digue, A. Damoy, B. Jeanjean, Overview of JOKER@CLEF 2022: Automatic wordplay and humour translation workshop, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, 2022, pp. 447–469.

[6] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the CLEF 2025 JOKER Task 1: Humour-aware Information Retrieval, in: [23], 2025.

[7] L. Ermakova, T. Miller, Y. Naud, A.-G. Bosser, R. Campos, Overview of the CLEF 2025 JOKER Task 3: Onomastic Wordplay Translation, in: [23], 2025.

[8] A. Kreefft-Libiu, F. Helms, C. Selçuk, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2025 JOKER Track, in: [23], 2025.

[9] G. Arampatzis, A. Arampatzis, DUTH at CLEF JOKER 2025 Tasks 2 and 3: Translating Puns and Proper Names with Neural Approaches, in: [23], 2025.

[10] S. K. P, B. A, S. M, T. S, REC_Cryptix at JOKER CLEF 2025: Teaching Machines to Laugh: Multilingual Humor Detection and Translation, in: [23], 2025.

[11] I. Kuzmin, CLEF 2025 JOKER track: No pun left behind, in: [23], 2025.

[12] P. Vachharajani, pjmathematician at the CLEF 2025 JOKER Lab Tasks 1, 2 & 3: A Unified Approach to Humour Retrieval and Translation using the Qwen LLM Family, in: [23], 2025.

[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: [23], 2025.

[14] D. A. M. Tobon, J. D. Jimenez, J. Serrano, J. C. M. Santos, E. Puertas, UTBNLP at CLEF JOKER 2025 Task 2: mBART-50 Fine-Tuning with Dictionary-Guided Forced Decoding and Phoneme-Based Techniques for English-French Pun Translation, in: [23], 2025.

[15] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022). doi:10.1016/j.patter.2022.100543.

[16] 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, 2002, pp. 311–318. URL: https://www.aclweb.org/anthology/P02-1040. doi:10.3115/1073083.1073135.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[18] L. Ermakova, T. Miller, A.-G. 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 CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1804–1817.

[19] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER Corpus: English–French parallel data for multilingual wordplay recognition, in: SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.

[20] L. Ermakova, A.-G. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER Task 3: Translate puns from English to French, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1800–1810.

[21] M. Post, A call for clarity in reporting BLEU scores, in: O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol,

M. Neves, M. Post, L. Specia, M. Turchi, K. Verspoor (Eds.), Proceedings of the Third Conference on Machine Translation: Research Papers, 2018, pp. 186–191. doi:10.18653/v1/W18-6319.

[22] H. Xu, K. Murray, P. Koehn, H. Hoang, A. Eriguchi, H. Khayrallah, X-alma: Plug & play modules and adaptive rejection for quality translation at scale, 2025. URL: https://arxiv.org/abs/2410.03115. arXiv:2410.03115.

[23] 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.

[24] L. Ermakova, T. Miller, A.-G. 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 CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1818–1827.

Table 7
Task 2 results in terms of the number of successful translations (manual evaluation)

run ID	count	# success	%
dsgt_o4_mini_multi_agent_discriminator	42	37	74
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	42	36	72
duth_xanthi_argos	50	26	52
duth_xanthi_GoogleTranslate_fallback	50	25	50
duth_google_flant5_fallback	50	25	50
Cryptix	50	25	50
duth_xanthi_GoogleTranslate	50	25	50
UvA_finetunedNLLB-1.3B	50	24	48
UvA_finetunedNLLB-1.3B&finetunedroBERTa	50	24	48
duth_xanthi_GoogleTranslate_fallback	48	24	48
pjmathematician_Q25-14	50	23	46
UvA_finetunedT5-base	50	23	46
UvA_T5-base&finetunedroBERTa	50	23	46
Skommarkhos_Lucie_SFT_ARPO	50	22	44
UvA_finetunedMarianMT	50	22	44
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	50	22	44
UvA_finetunedMarianMT&finetunedroBERTa	50	22	44
pjmathematician_Q25-14	50	21	42
teamX_aug	50	21	42
Skommarkhos_Lucie_SFT	50	21	42
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	50	21	42
teamX_aug	50	20	40
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	50	20	40
dsgt_simple_mistral_medium	42	20	40
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	50	20	40
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	50	20	40
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	50	20	40
Skommarkhos_Lucie-7B-Instruct-v1.1	50	20	40
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	50	20	40
Cryptix	50	19	38
yourteamid_marianmt_pun_postedit	50	19	38
duth_xanthi_helsinki	50	19	38
Cryptix_marianmt	50	19	38
duth_hybrid_fusion	50	19	38
Skommarkhos_Croissant_SFT_ARPO	50	18	36
duth_xanthi_m2m100_1_2B	50	18	36
Skommarkhos_Croissant_SFT	50	17	34
duth_xanthi_t5_base_gpu	50	17	34
duth_combined_m2m100	50	16	32
UvA_finetunedmBARTcc25	50	15	30
duth_xanthi_mbart50	50	14	28
teamX_final	50	13	26
dsgt_simple_o4_mini	42	12	24
duth_xanthi_bloomz3b_local	50	11	22
UvA_mBARTcc25&finetunedroBERTa	50	11	22
Cryptix_finetunedmarian	49	2	4
yourteam_rulebased	3	1	2
Cryptix_rulebased	3	1	2

Table 8
Task 2 results in terms of BLEU score and BLEU n-gram precision (training data)

run ID	score	$n=1$	$n=2$	$n=3$	$n=4$
Cryptix_rulebased	100.00	100.00	100.00	100.00	100.00
pjmathematician_Q25-14	100.00	100.00	100.00	100.00	100.00
yourteam_rulebased	100.00	100.00	100.00	100.00	100.00
teamX_final	99.77	99.83	99.79	99.76	99.75
UvA_finetunedNLLB-1.3B	54.08	73.99	57.16	48.42	41.77
UvA_finetunedNLLB-1.3B&finetunedroBERTa	53.66	73.57	56.75	48.03	41.34
UvA_finetunedMarianMT	53.38	73.73	56.49	47.55	40.98
UvA_finetunedMarianMT&finetunedroBERTa	52.88	73.28	56.07	47.08	40.43
Cryptix_marianmt	50.09	71.47	52.93	43.97	38.02
Cryptix	50.05	71.52	52.97	44.03	38.08
yourteamid_marianmt_pun_postedit	50.05	71.52	52.97	44.03	38.08
duth_xanthi_helsinki	50.05	71.52	52.97	44.03	38.08
duth_hybrid_fusion	50.02	71.49	52.93	43.98	38.04
Skommarkhos_Lucie_SFT	49.98	72.04	53.64	43.95	36.73
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	48.73	71.32	52.46	42.57	35.41
duth_xanthi_argos	47.92	70.75	51.70	42.33	35.73
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	47.88	70.80	51.70	41.68	34.45
Skommarkhos_Lucie_SFT_ARPO	46.44	68.89	50.12	40.44	33.32
Skommarkhos_Lucie-7B-Instruct-v1.1	45.92	68.25	49.52	39.93	32.96
Skommarkhos_Lucie-7B-Instruct-v1.1	45.75	68.10	49.35	39.76	32.79
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	45.65	68.03	49.25	39.64	32.69
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	45.63	68.12	49.23	39.62	32.64
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	45.56	68.01	49.18	39.58	32.55
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	45.55	67.99	49.17	39.57	32.56
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	45.50	67.90	49.12	39.50	32.53
UvA_finetunedT5-base	44.93	68.76	48.88	38.65	31.37
UvA_T5-base&finetunedroBERTa	44.22	68.23	48.13	37.94	30.70
Cryptix	43.84	68.61	47.88	37.50	29.99
duth_xanthi_GoogleTranslate_fallback	43.81	68.63	47.86	37.45	29.93
duth_xanthi_GoogleTranslate	43.75	68.57	47.80	37.41	29.89
duth_xanthi_GoogleTranslate_fallback	43.35	68.13	47.32	37.02	29.59
teamX_aug	41.80	61.73	44.87	36.34	30.33
duth_google_flant5_fallback	41.71	66.44	45.60	35.43	28.19
duth_xanthi_m2m100_1_2B	41.07	67.43	45.91	35.12	27.63
Skommarkhos_Croissant_SFT	40.49	66.05	44.78	34.09	26.65
Skommarkhos_Croissant_SFT_ARPO	40.04	65.81	44.38	33.64	26.15
duth_xanthi_t5_base_gpu	38.47	63.59	42.37	32.20	25.23
duth_xanthi_mbart50	35.87	63.41	40.41	29.49	21.92
teamX_aug	35.56	64.19	40.61	28.91	21.22
duth_combined_m2m100	34.91	63.13	39.85	28.82	21.42
UvA_mBARTcc25&finetunedroBERTa	33.58	56.58	36.79	28.00	21.82
UvA_finetunedmBARTcc25	28.46	52.73	31.79	22.97	17.04
dsgt_o4_mini_multi_agent_discriminator	24.16	52.00	28.08	18.52	12.60
duth_xanthi_bloomz3b_local	20.05	47.55	23.17	14.82	9.90
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	19.30	46.43	22.89	14.22	9.19
dsgt_simple_mistral_medium	17.34	44.36	20.34	12.55	7.99
dsgt_simple_o4_mini	9.87	35.44	11.67	6.27	3.65
Cryptix_finetunedmarian	0.34	14.32	0.87	0.09	0.01

Table 9
Task 2 results in terms of BERTScore precision, recall, and F_1 (training data)

run ID	P	R	F_1
teamX_final	83.81	84.37	84.07
pjmathematician_Q25-14	83.97	83.88	83.90
yourteam_rulebased	83.97	83.88	83.90
Cryptix_rulebased	83.97	83.88	83.90
UvA_finetunedNLLB-1.3B	83.75	82.91	83.30
UvA_finetunedNLLB-1.3B&finetunedroBERTa	83.58	82.89	83.20
UvA_finetunedMarianMT	83.51	82.53	82.99
UvA_finetunedMarianMT&finetunedroBERTa	83.38	82.52	82.92
Skommarkhos_Lucie_SFT	83.32	82.54	82.90
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	82.74	82.04	82.36
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	82.53	81.97	82.22
UvA_finetunedT5-base	82.28	81.65	81.93
UvA_T5-base&finetunedroBERTa	82.14	81.62	81.85
Cryptix	82.03	81.33	81.65
duth_xanthi_GoogleTranslate	82.02	81.31	81.63
duth_xanthi_GoogleTranslate_fallback	82.02	81.30	81.63
Cryptix_marianmt	82.02	81.17	81.55
yourteamid_marianmt_pun_postedit	82.02	81.16	81.54
duth_xanthi_helsinki	82.02	81.16	81.54
Cryptix	82.02	81.16	81.54
duth_hybrid_fusion	82.02	81.16	81.54
duth_xanthi_GoogleTranslate_fallback	81.74	81.03	81.35
teamX_aug	81.27	81.26	81.17
Skommarkhos_Croissant_SFT	81.34	81.06	81.16
duth_xanthi_argos	81.60	80.73	81.12
Skommarkhos_Croissant_SFT_ARPO	81.27	80.99	81.09
Skommarkhos_Lucie_SFT_ARPO	81.05	80.98	80.97
duth_xanthi_m2m100_1_2B	81.30	80.66	80.94
duth_google_flant5_fallback	81.03	80.42	80.69
Skommarkhos_Lucie-7B-Instruct-v1.1	80.62	80.64	80.59
teamX_aug	80.81	80.28	80.51
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	80.56	80.55	80.51
Skommarkhos_Lucie-7B-Instruct-v1.1	80.52	80.55	80.49
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	80.54	80.50	80.48
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	80.52	80.49	80.46
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	80.49	80.47	80.44
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	80.40	80.42	80.36
duth_xanthi_mbart50	80.31	79.78	80.01
duth_combined_m2m100	80.18	79.43	79.77
UvA_mBARTcc25&finetunedroBERTa	79.90	79.52	79.68
duth_xanthi_t5_base_gpu	79.25	79.35	79.26
UvA_finetunedmBARTcc25	78.89	78.68	78.74
dsgt_o4_mini_multi_agent_discriminator	77.28	78.60	77.89
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	75.70	76.96	76.28
duth_xanthi_bloomz3b_local	76.57	76.01	76.25
dsgt_simple_mistral_medium	75.90	76.65	76.23
dsgt_simple_o4_mini	73.06	73.16	73.08
Cryptix_finetunedmarian	72.76	71.42	72.05

Table 10
Task 2 results in terms of the pun location–based metric (training data)

run ID	count	location	%
Cryptix_rulebased	1405	391	27.83
pjmathematician_Q25-14	1405	391	27.83
yourteam_rulebased	1405	391	27.83
teamX_final	1405	380	27.05
dsgt_o4_mini_chain_of_thought_phonetic_embeddings	1405	184	13.10
dsgt_o4_mini_multi_agent_discriminator	1405	183	13.02
UvA_finetunedMarianMT&finetunedroBERTa	1405	178	12.67
UvA_finetunedMarianMT	1405	172	12.24
UvA_finetunedNLLB-1.3B	1405	169	12.03
UvA_finetunedNLLB-1.3B&finetunedroBERTa	1405	169	12.03
teamX_aug	1405	166	11.81
Cryptix_marianmt	1405	162	11.53
duth_hybrid_fusion	1405	161	11.46
Cryptix	1405	161	11.46
yourteamid_marianmt_pun_postedit	1405	161	11.46
duth_xanthi_helsinki	1405	161	11.46
duth_xanthi_argos	1405	154	10.96
UvA_mBARTcc25&finetunedroBERTa	1405	153	10.89
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a5	1405	150	10.68
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a7	1405	149	10.60
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_arpo_a19	1405	149	10.60
Skommarkhos_Lucie-7B-Instruct-v1.1	1405	149	10.60
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a11	1405	148	10.53
Skommarkhos_Lucie-7B-Instruct-v1.1	1405	148	10.53
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v8	1405	146	10.39
Cryptix	1405	145	10.32
UvA_finetunedT5-base	1405	144	10.25
Skommarkhos_skommarkhos_lucie7binstructv1_1_sft_v4	1405	144	10.25
duth_xanthi_GoogleTranslate	1405	143	10.18
duth_xanthi_GoogleTranslate_fallback	1405	143	10.18
Skommarkhos_skommarkhos-lucie7binstructv1-1-sft-arpo-a1	1405	143	10.18
Skommarkhos_Lucie_SFT_ARPO	1405	142	10.11
duth_xanthi_t5_base_gpu	1405	140	9.96
duth_google_flant5_fallback	1405	140	9.96
UvA_T5-base&finetunedroBERTa	1405	140	9.96
Skommarkhos_Lucie_SFT	1405	137	9.75
duth_xanthi_m2m100_1_2B	1405	135	9.61
UvA_finetunedmBARTcc25	1405	132	9.40
Skommarkhos_Croissant_SFT_ARPO	1405	131	9.32
Skommarkhos_Croissant_SFT	1405	126	8.97
teamX_aug	1405	119	8.47
duth_xanthi_mbart50	1405	118	8.40
duth_combined_m2m100	1405	113	8.04
duth_xanthi_bloomz3b_local	1405	76	5.41
dsgt_simple_mistral_medium	1405	58	4.13
Cryptix_finetunedmarian	1405	35	2.49
dsgt_simple_o4_mini	1405	25	1.78