Overview of JOKER 2023 Automatic Wordplay Analysis Task 3 – Pun translation

Liana Ermakova[1,*], Tristan Miller[2], Anne-Gwenn Bosser[3], Victor Manuel Palma Preciado[1,4], Grigori Sidorov[4] and Adam Jatowt[5]

[1] *Université de Bretagne Occidentale, HCTI, France*

[2] *Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria*

[3] *École Nationale d'Ingénieurs de Brest, Lab-STICC CNRS UMR 6285, France*

[4] *Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*

[4] *University of Innsbruck, Austria*

Abstract

This paper provides a comprehensive overview of Task 3 of the JOKER-2023 track. The overarching objective of the JOKER track series is to facilitate collaboration among linguists, translators, and computer scientists to advance the development of automatic interpretation, generation, and translation of wordplay. Task 3 specifically concentrates on the automatic translation of puns from English into French and Spanish. In this overview, we outline the overall structure of the shared task that we organized as part of the CLEF-2023 evaluation campaign. We discuss the approaches employed by the participants and present and analyze the results they achieved.

Keywords

wordplay, puns, computational humour, machine translation

1. Introduction

This paper describes Task 3 of the JOKER-2023[1] challenge, where the goal is to accurately translate puns between different languages. This is the final task of JOKER-2023 [1], following Tasks 1 [2] and 2 [3] on pun detection and pun location/slash interpretation, respectively.

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 3, participating systems attempt to translate English punning jokes into French and Spanish. The translations should aim to preserve, to the extent possible, both the form and meaning of the original wordplay – that is, to implement the pun→pun strategy described in Delabastita's typology of pun translation strategies [4, 5]. For example, "I used to be a banker but I lost interest" might be rendered into French as "*J'ai été*

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

0000-0002-7598-7474 (L. Ermakova); 0000-0002-0749-1100 (T. Miller); 0000-0002-0442-2660 (A. Bosser); 0000-0001-8711-1106 (V. M. Palma Preciado); 0000-0003-3901-3522 (G. Sidorov); 0000-0001-7235-0665 (A. Jatowt)

CEUR Workshop Proceedings (CEUR-WS.org)

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

Table 1
Task 3 dataset statistics

	Train		Test	
Language	target	source	target	source
French	5,838	1,405	6,590	1,197
Spanish	644	217	5,727	544

banquier mais j'en ai perdu tout l'intérêt". This fairly straightforward translation preserves the pun, since *interest* and *intérêt* share the same semantic ambiguity.

In the remainder of this paper, we describe the data preparation process (Section 2) and participants' approaches (Section 3), and then present an analysis of their results (Section 4). Section 5 concludes the paper.

2. Data

Our French training data contains 5,838 translations of 1,405 distinct puns in English as used in Tasks 1 and 2. These translations come from translation contests and the JOKER-2022 track [6, 7]. For the test set, we provided participants with 4,290 distinct puns in English to be translated into French and Spanish. A detailed description of the corpus can be found in our SIGIR 2023 paper [8].

We also provide new sets of English–Spanish translations of punning jokes, similar to the English–French datasets we produced for JOKER-2022. These translations were sourced via a translation contest in which professional translators were asked to translate 400 English puns. In total, they produced 2,459 pairs of translated puns. These translations underwent an expert review to ensure compliance with the data set's criteria of preserving both wordplay and the general meaning. We kept 644 translations of 217 distinct English puns for training data.

Statistics on the dataset are given in Table 1. As in cases of Tasks 1 and 2, we included the training data in the input file of the test data. This allows us for comparison of the systems both on the test and training sets.

As described below, the data was provided in JSON and delimited text formats with fields containing the text of the punning joke and a unique ID; for training there were one or two additional fields containing gold-standard translations of the text into French and/or Spanish. Systems were expected to output a JSON or delimited text file containing the run ID, text ID, the text of the translation(s) into French and/or Spanish, and a boolean flag indicating whether the run was manual or automatic.

Input format. The base data is provided in JSON and CSV formats with the following fields:

id_en a unique identifier

text_en the text of the instance of source wordplay in English

Input example:

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

Qrels. We provide training data as JSON or TSV qrels files with the following fields:

id_en a unique identifier from the input file

text_fr (optional) translation of the wordplay into French

text_es (optional) translation of the wordplay into Spanish

Example of a qrel file:

```
[{"id_en":"en_1",
"text_fr":"J'ai été banquier mais j'en ai perdu tout l'intérêt"}]
```

Output Format. Participating systems were expected to submit their results as a TREC-style JSON or TSV file with the following fields:

run_id run ID starting with <team_id>_<task_id>_<method_used> – e.g., UBO_BLOOM

manual whether the run is manual (0 or 1)

id_en a unique identifier from the input file

text_fr (optional) translation of the wordplay into French

text_es (optional) translation of the wordplay into Spanish

Example of an output file:

```
[{"run_id":"team1_task_3_DeepL",
"manual":0,
"id_en":"en_1",
"text_fr":"J'ai été banquier mais j'en ai perdu tout l'intérêt"}
]
```

3. Participants' approaches

Nine teams submitted 47 runs for this task, as summarized in Table 2. The approaches used were as follows:

1. The LJGG team submitted runs for translation from English to French and Spanish. Their model is a three-stage architecture based on T5 (SimpleT5). The two stages calculate the information necessary to concatenate the English sentence, which forms an input for the third neural network. For training the models, they enlarged Task 3's dataset with the data prepared for Task 1. They also used the DeepL translator to compare their results and found that the DeepL translations are better.

Table 2
Statistics on the runs submitted for Task 3

Team	EN→FR	EN→ES
Croland	1	1
LJGG	4	5
MiCroGerk	—	7
Smroltra	6	6
TeamCAU	3	—
TheLangVerse	1	1
ThePunDetectives	2	2
UBO	3	3
NPalma	—	2
Total	20	27

2. The NLPalma team [9] approached the translation of wordplay from English to Spanish using BLOOMZ & mT5, which is an improved version of BLOOM.
3. The MiCroGerk team [10] used SimpleT5-, BLOOM-, OpenAI-, and AI21-based models and the models from the EasyNMT package (Opus-MT, mBART50_m2m, and M2M_10) for the English–Spanish translation task. The OpenAI- and AI21-based models proved to be the best, with the lowest-ranked models being SimpleT5. According to the authors, however, there is still plenty of room for improvement.
4. The UBO team [11] used the models from the EasyNMT package – namely, Opus-MT, mBART50_m2m, and M2M_100.
5. The TheLangVerse team made use of the j2-grande model from the AI21 platform. They also combined the datasets to provide more content for fine-tuning, obtaining results comparable to those obtained from their surveys.
6. Opus-MT and M2M_100 from the the EasyNMT package were selected by participants of ThePunDetectives team [12]. The authors found that M2M_100 made translations that diverged from the original senses at the expense of precision. In contrast, Opus-MT presented a slightly better translation capability, being able to comprehend some types of humour.
7. The solution of the Smroltra team [13] was to use the GPT-3, BLOOM, Opus-MT, and mBART50_m2m models from EasyNMT; SimpleT5; and the Google Translate service for both English–Spanish and English–French translations. The best results were obtained using GPT-3, while the worst came from T5, which produced incoherent sentences. GPT-3 and BLOOM obtained the highest scores on both datasets, although according to the authors, the translation of the datasets requires more data and time.
8. The Croland team [14] approached the task using GPT-3.
9. TeamCAU [15] report using large language models (LLMs), but do not specifically describe their use for their Task 3 runs.

4. Results

We continue JOKER-2022's practice of having trained experts manually evaluate system translations according to features such as sense preservation and wordplay, since vocabulary overlap metrics such as BLEU are unsuitable for evaluating wordplay translations [7, 6]. Participants' runs were subject to whitespace trimming and lower-casing, and were pooled together. We then filtered out French and Spanish translations identical to the original wordplay in English, as we considered these wordplay instances to be untranslated. Then, we manually evaluated 6,590 French translations of 1,197 distinct puns in English pooled from the participants' runs used as the final test data. Besides, our experts manually assessed 9,682 French translations of 868 distinct puns in English. We manually evaluated 5,727 Spanish translations of 544 distinct English puns. The 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.

Table 3 shows the results on the test data while Table 4 displays the results obtained on the training data for the pun translation task from English into French. The following scores are reported in both tables:

#E number of manually evaluated translations

#T number of submitted translations used for evaluation

#M number of translations preserving the meaning of the source puns

%M percentage of translations preserving the meaning of the source puns

#W number of translations containing wordplay

%W percentage of translations containing wordplay

#S number of translations containing wordplay and preserving the meaning of the source puns

%S percentage of translations containing wordplay and preserving the meaning of the source puns

%R percentage of translations containing wordplay and preserving the meaning of the source puns over the total test set

We will consider **#S** measure as the one for ranking the submitted runs. We observe that for English to French translation, the Jurassic-2 and T5 models obtained the best results (respectively, 72 and 65 translations that contain the wordplay and preserve the meaning of the source puns). We should note here, however, that the T5 model was trained on the training set while other LLMs were used only in a few-shot setup. Overall, same as in 2022 [7, 6], we notice that the success rate of wordplay translation is very low, and the task is obviously very challenging. This is even the case for LLMs, with a maximum value of 6% over the total evaluated test set for French. The results are almost three times higher for the training set in French, suggesting an overfitting problem but still very low in general (less than 17%)

For English-to-Spanish translation, the best results were obtained by systems that used the Google Translate service (96 or 99 correctly translated puns) and ones based on the mBART

Table 3
Results for pun translation from English into French (test data)

run ID	#E	#T	#M	%M	#W	%W	#S	%S	%R
Croland_task_3_EN_FR_GPT3	16	28	4	25	0	0	0	0	0
LJGG_Google_Translator_EN_FR_auto	1,076	1,197	580	53	67	6	63	5	5
LJGG_task3_fr_mt5_base_auto	2	1,197	2	100	1	50	1	50	0
LJGG_task3_fr_mt5_base_no_label_auto	1	1,197	1	100	0	0	0	0	0
LJGG_task3_fr_t5_large_auto	90	1,197	24	26	2	2	2	2	0
LJGG_task3_fr_t5_large_no_label_auto	140	1,197	80	57	15	10	15	10	1
Smroltra_task_3_EN-FR_BLOOM	31	32	8	25	0	0	0	0	0
Smroltra_task_3_EN-FR_EasyNMT-Opus	786	1,197	427	54	58	7	56	7	4
Smroltra_task_3_EN-FR_EasyNMT-mbart	1,139	1,197	613	53	68	5	64	5	5
Smroltra_task_3_EN-FR_GPT3	30	32	8	26	0	0	0	0	0
Smroltra_task_3_EN-FR_GoogleTranslation	1,109	1,197	602	54	71	6	67	6	5
Smroltra_task_3_EN-FR_SimpleT5	1,043	1,197	562	53	66	6	65	6	5
TeamCAU_task_3_EN-FR_AI21	30	32	8	26	0	0	0	0	0
TeamCAU_task_3_EN-FR_BLOOM	32	32	8	25	0	0	0	0	0
TeamCAU_task_3_EN-FR_ST5	1,090	1,197	577	52	71	6	69	6	5
TheLangVerse_task_3_j2-grande-finetuned	1,176	1,197	636	54	76	6	**72**	6	6
ThePunDetectives_task_3_EN-FR_M2M100	13	340	9	69	2	15	2	15	0
ThePunDetectives_task_3_EN-FR_OpusMT	183	340	92	50	19	10	17	9	1
UBO_task_3_SimpleT5	73	1,195	47	64	5	6	5	6	0
UBO_task_3_SimpleT5_x	1,148	1,195	616	53	71	6	67	5	5
UBO_task_3_SimpleT5_y	791	1,194	429	54	61	7	59	7	5

model (99 puns). The maximum score is 18% in the case of English-to-Spanish translation, which is considerably higher than for French data. Note that for the Spanish version of the task we only have evaluations on the test data.

5. Conclusion

In this paper, we have described Task 3 of the JOKER track at CLEF 2023. The task aims to advance the automation of wordplay translation, and included shared tasks on translation from English to French and from English to Spanish. We expanded the EN→FR training set described in our SIGIR 2023 paper [8] with a new parallel corpus of EN→ES wordplay translations. We evaluated the results from participants after pooling and conducting manual assessments with experts.

We observe that the success rate of wordplay translation is extremely low even in the case of LLMs, for both language pairs. The maximum value of 6% over the total evaluated test set

Table 4
Results for pun translation from English into French (training data)

run ID	#E	#T	#M	%M	#W	%W	#S	%S	%R
Croland_task_3_EN_FR_GPT3	17	32	8	47	0	0	0	0	0
LJGG_Google_Translator_EN_FR_auto	757	868	451	60	128	17	124	16	14
LJGG_task3_fr_t5_large_auto	21	868	3	14	1	5	1	5	0
LJGG_task3_fr_t5_large_no_label_auto	88	868	54	61	26	30	22	25	3
Smroltra_task_3_EN-FR_BLOOM	31	36	11	35	0	0	0	0	0
Smroltra_task_3_EN-FR_EasyNMT-Opus	432	868	250	58	65	15	64	15	7
Smroltra_task_3_EN-FR_EasyNMT-mbart	793	868	470	59	143	18	136	17	16
Smroltra_task_3_EN-FR_GPT3	30	36	13	43	0	0	0	0	0
Smroltra_task_3_EN-FR_GoogleTranslation	746	868	444	60	126	17	122	16	14
Smroltra_task_3_EN-FR_SimpleT5	697	868	412	59	105	15	100	14	12
TeamCAU_task_3_EN-FR_AI21	32	36	13	41	0	0	0	0	0
TeamCAU_task_3_EN-FR_BLOOM	29	36	12	41	0	0	0	0	0
TeamCAU_task_3_EN-FR_ST5	683	868	405	59	97	14	92	13	11
TheLangVerse_task_3_j2-grande-finetuned	675	868	405	60	127	19	122	18	14
ThePunDetectives_task_3_EN-FR_M2M100	22	321	16	73	9	41	9	41	1
ThePunDetectives_task_3_EN-FR_OpusMT	164	321	95	58	24	15	24	15	3
UBO_task_3_SimpleT5	38	868	28	74	15	39	15	39	2
UBO_task_3_SimpleT5_x	810	868	486	60	148	18	141	17	16
UBO_task_3_SimpleT5_y	442	868	255	58	66	15	65	15	7

was obtained for French while the corresponding value for Spanish was 18%. In French, even for the training set the percentage of successful translations is less than 17%. The difficulty of translation of wordplay between relatively well-studied languages, even when using LLMs, calls for more community attention to this challenging task. Among the submitted runs, those using mBART, Jurassic-2, T5, and Google Translate produced the best results. As with Tasks 1 and 2, we received many partial runs due to the constraints involved in using LLMs.

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

Acknowledgments

This project 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. JOKER is supported by *La Maison des sciences de l'homme en Bretagne*. For their help and support in the first Spanish pun translation contest, we thank Carolina Palma Preciado, Leopoldo Jesús Gutierrez Galeano, Khatima El Krirh, Nathalie Narváez Bruneau, and Rachel

Table 5
Results for pun translation from English into Spanish (test data)

run ID	#E	#T	#M	%M	#W	%W	#S	%S	%R
Croland_task_3_ENESGPT3	45	47	9	20.00	3	6.66	3	6.66	0
LJGG_task3_es_mt5_base_auto	34	544	16	47.05	5	14.70	5	14.70	0
LJGG_task3_es_mt5_base_no_label_auto	34	544	16	47.05	5	14.70	5	14.70	0
LJGG_task3_es_t5_large_auto	34	544	16	47.05	5	14.70	5	14.70	0
LJGG_task3_es_t5_large_no_label_auto	34	544	16	47.05	5	14.70	5	14.70	0
LJGG_task_3_GoogleTranslatorENESauto	544	544	274	50.36	106	19.48	**99**	18.19	18
NLPalma_task_3_BLOOMZ_x	359	359	215	59.88	85	23.67	80	22.28	14
NLPalma_task_3_BLOOMZ_y	359	359	215	59.88	85	23.67	80	22.28	14
Smroltra_task_3_EN-ES_EasyNMT-Opus	529	544	263	49.71	100	18.90	93	17.58	17
Smroltra_task_3_EN-ES_EasyNMT-Opus_x	529	544	263	49.71	100	18.90	93	17.58	17
Smroltra_task_3_EN-ES_EasyNMT-Opus_y	529	544	263	49.71	100	18.90	93	17.58	17
Smroltra_task_3_EN-ES_GoogleTranslation	532	544	267	50.18	103	19.36	96	18.04	17
Smroltra_task_3_EN-ES_SimpleT5	531	544	265	49.90	101	19.02	94	17.70	17
Smroltra_task_3_ENESBLOOM	45	47	8	17.77	2	4.44	2	4.44	0
TheLangVerse_task_3_j2-grande-finetuned	415	544	200	48.19	70	16.86	65	15.66	11
ThePunDetectives_task_3_EN-ES_M2M100	33	430	16	48.48	7	21.21	7	21.21	1
ThePunDetectives_task_3_ENESOpusMT	428	430	208	48.59	71	16.58	66	15.42	12
MiCroGerk_task_3_EN-ES_OpenAI	6	17	3	0.5	1	16.66	1	16.66	0
MiCroGerk_task_3_EN-ES_mbart50_m2m_x	543	544	274	50.46	106	19.52	99	18.23	18
MiCroGerk_task_3_EN-ES_AI21_x	1	17	1	1	0	0	0	0	0
MiCroGerk_task_3_EN-ES_mbart50_m2m_y	543	544	274	50.46	106	19.52	99	18.23	18
MiCroGerk_task_3_EN-ES_m2m_100_418M	43	544	23	53.48	11	25.58	11	25.58	2
MiCroGerk_task_3_EN-ES_SimpleT5	5	544	4	0.8	3	0.6	3	0.6	0

Kinlay. We also thank all other colleagues and students who participated in data construction, the translation contests, and the CLEF JOKER track.

References

[1] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. 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.), CLEF'23: Proceedings of the Fourteenth International Confer-

ence of the CLEF Association, Lecture Notes in Computer Science, Springer, 2023.

[2] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 1 - pun detection, in: [16], 2023.

[3] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 2 - Pun location and interpretation, in: [16], 2023.

[4] D. Delabastita, There's a Double Tongue: an Investigation into the Translation of Shakespeare's Wordplay, with Special Reference to Hamlet, Rodopi, Amsterdam, 1993.

[5] D. Delabastita, Wordplay as a translation problem: a linguistic perspective, in: Ein internationales Handbuch zur Übersetzungsforschung, volume 1, De Gruyter Mouton, 2008, pp. 600–606. doi:10.1515/9783110137088.1.6.600.

[6] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, C. Borg, Élise 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: Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, Cham, 2022, pp. 447–469. doi:10.1007/978-3-031-13643-6_27.

[7] L. Ermakova, F. Regattin, T. Miller, A.-G. Bosser, C. Borg, B. Jeanjean, Élise 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*, 2022, pp. 1681–1700.

[8] 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. doi:10.1145/3539618.3591885, to appear.

[9] V. M. P. Preciado, C. P. Preciado, G. Sidorov, NLPalma @ CLEF 2023 JOKER: A BLOOMZ and BERT approach for wordplay detection and translation, in: [16], 2023.

[10] A. Prnjak, D. R. Davari, K. Schmitt, CLEF 2023 JOKER Task 1, 2, 3: pun detection, pun interpretation, and pun translation, in: [16], 2023.

[11] Q. Dubreuil, UBO Team @ CLEF JOKER 2023 track for Task 1, 2 and 3 - applying AI models in regards to pun translation, in: [16], 2023.

[12] F. Ohnesorge, M. Á. Gutiérrez, J. Plichta, CLEF 2023 JOKER Tasks 2 and 3: using NLP models for pun location, interpretation and translation, in: [16], 2023.

[13] O. Popova, P. Dadić, Does AI have a sense of humor? CLEF 2023 JOKER tasks 1, 2 and 3: Using BLOOM, GPT, SimpleT5, and more for pun detection, location, interpretation and translation, in: [16], 2023.

[14] J. Komorowska, I. Čatipović, D. Vujica, CLEF2023' JOKER Working Notes, in: [16], 2023.

[15] A. Anjum, N. Lieberum, Exploring Humor in Natural Language Processing: A Comprehensive Review of JOKER Tasks at CLEF Symposium 2023, in: [16], 2023.

[16] Proceedings of the Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2023.