Overview of the CLEF 2025 JOKER Task 3: Onomastic Wordplay Translation

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

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

Abstract

This paper summarises the setup and results of the shared task on onomastic wordplay translation at the CLEF 2025 JOKER Lab, an earlier version of which was run at JOKER 2022 as a pilot task. The objective of the task is to translate wordplay concerned with proper names from English to French. Such wordplay, widespread in classic and modern creative writing, is particularly challenging to translate due to its idiosyncratic nature and cultural references. Four teams participated in this year's task, submitting 20 runs. We describe our construction of the data set using for training and testing, the methods employed by the participating teams, and the results obtained for the runs and a naïve baseline in terms of various manually and automatically applied measures of translation quality. Despite notable advances, we find that translation of onomastic wordplay remains highly challenging, with fewer than 10% of manually evaluated translations judged as acceptable alternatives. Recurrent errors included untranslated source wordplay, overfitting to the training data, omission of surnames, and nonsensical generations.

Keywords

wordplay, computation humour, named entities, neologisms, machine translation, LLM, transformers

1. Introduction

This paper describes Task 3 of the JOKER-2025 Track [1], which aims to benchmark automatic translation of onomastic (i.e., name-related) wordplay from English to French. A pilot version of the task, on machine translation of wordplay in named entities was, run in the JOKER's 2022 edition [2, 3] and employed a parallel corpus of onomastic wordplay in English and French that we constructed from video games, advertising slogans, literature, and other sources. This year we extended the corpus with new onomastic wordplay instances; we also provided short contexts for the names, which are often necessary to recognise, understand, and translate the wordplay they contain. This year, Task 3 complements two other tasks, one on humour-aware information retrieval [4] and the other on translation of English puns into French [5].

Wordplay is often used for its attention-getting or mnemonic qualities in headlines, toponyms, company names, and advertising. Onomastic wordplay is used as a rhetorical device by novelists, poets and playwrights. It is widespread in classic literature [6], such as in Shakespeare's characters' names [7], but also in names found in modern-day works such as Pokémon, Harry Potter, Asterix, and video games. Proper nouns with an extra semantic load are used as a meaningful element in literary texts and can be considered as wordplay [8]. The translation of such names is problematic, raising the questions of whether the transposition of such names into a given target language is technically possible and, if so, what method might be appropriate for doing this [8]. Common approaches to translating names include

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

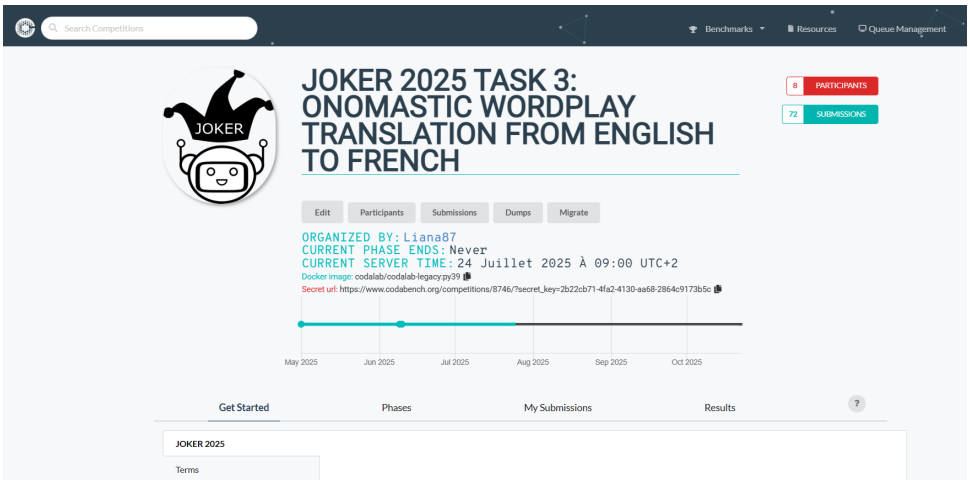

Figure 1: CLEF 2025 JOKER Task 3 on Codabench

Table 1
Number of runs submitted to Task 3

team	#
arampageos [12]	15
mariapazr20 [13]	1
pjmathematician [14]	3
sarath_kumar [15]	1
Total	22

transliterating them [9, 10] or keeping them unchanged in the target text. However, these approaches rarely preserve wordplay in a meaningful way, which may harm the text's pragmatic force.

This year, the JOKER track ran its shared tasks through Codabench[1] [11], a Free Software online platform for organising AI benchmarks. (See Figure 1.) This greatly facilitated running shared tasks and attracted many new participants. Although we are continuing to receive new registrations and post-competition submissions, this paper presents only those runs submitted before our official results were communicated to the participants. As summarised in Table 1, four teams participated in JOKER Task 3, submitting a total of 20 official runs via the Codabench platform.

In the remainder of this paper, we present related work (Section 2), the data (Section 3), the evaluation measures (Section 4), participants' approaches (Section 5), and an analysis of their results for both training and test data (Section 6). Section 7 concludes the paper.

2. Related work

As natural language processing continues to develop, its relationship with translation studies remains an active – and debated – area of research. Studies highlight both the opportunities and challenges that AI faces in translation, the new possibilities that AI brings, as well as the issues posed by the subtlety of language, especially in humorous translation.

The combination of humour studies and translation has become increasingly relevant, particularly with the accelerating progress of AI technologies such as GPT models. The translation of humour – which frequently relies on linguistic nuance, cultural context, and wordplay – poses significant challenges to both human translators and AI models. These challenges are compounded when humour involves neologisms, as newly coined terms often rely on specific cultural or temporal contexts that may not yet

[1] https://www.codabench.org/competitions/8746/

have direct equivalents in the target language. Research in this field aims to identify these difficulties by studying how traditional translation models often struggle to preserve the intended humour of the source text in the target language.

Theoretical models of translation differ in terms of their relevance and sensitivities to humour, including their recognition of humour, their treatment of its unique features, and how they identify and solve various humour-related translation problems [16]. These differences are highly relevant for the case of neologisms, which are embedded in humorous contexts and as such may require both semantic interpretation and creative adaptation for an accurate translation.

The first challenge related to humour detection is the requirement to understand subtleties such as irony, tone, register, and various physical cues that contribute to humour [17]. This challenge is even greater when humour is expressed through wordplay and neologism, which are common in humorous writing and often cannot be expressed through a word-for-word translation.

The second challenge is the means by which translators navigate wordplay, cultural references, and issues related to censorship, sensitivity, and agency. For instance, translating humour that depends on archaic language or specific dialects can be highly challenging; however, there are indications that humour can be accurately translated by adopting a "bidirectional" sense of humour, which focuses on interpretation and re-creation rather than seeking an exact match in the target language [17]. This approach facilitates a more complex and creative translation of humour, and is particularly useful in the translation of neologisms, which often carry not just linguistic innovation but cultural commentary as well.

The emergence of AI as a publicly accessible technology – especially with models like those of ChatGPT, Claude, Llama or Gemini – has significantly impacted translation studies, including humour translation. A systematic review of research on GPT-based translation reveals that AI-generated translations often match human translations and can even surpass traditional machine translation in the treatment of complex language such as humour, wordplay, and even poetry [18]. However, as a study [19] of humorous translations generated by GPT-4 points out, AI still struggles to grasp the full depth of humour, particularly with respect to cultural context and subtlety. Although AI translations can provide comparable or even improved experiences in certain situations, there is manifestly room for improvement. This likely applies to neologisms, which have no standard definitions and require a flexible, context-sensitive approach that AI has yet to fully master.

The use of AI in humorous translation goes beyond traditional written texts. The use of AI in creative writing – especially in comedy – has been examined through work [e.g., 20] that views AI tools as collaborators for writers, not competitors. This cooperative approach creates new opportunities for humorous literature writers, allowing them to engage with AI as a 'new toy' in the creative process. Exploring AI's ability to generate and adapt neologisms could be particularly valuable in this process. Giving writers new linguistic tools to experiment with naturally raises broader questions about the quality of AI-generated content, but also helps reframe AI as a tool for creation rather than a threat to human creativity.

3. Dataset

3.1. Construction

Task 3 uses a parallel corpus of wordplay in named entities in English and French, drawn from video games, advertising slogans, literature, and other sources. This corpus is based in part on one used for the 2022 edition of this task; for that task we had sourced 1,398 names in English along with 1,450 translations into French. The vast majority of those translations are the official, published ones, and as such may already be included in the training data of popular large language models (LLMs). Some alternative translations had been provided by Master's students in translation, all native speakers of French. For some sources, such as *Pokémon* names, we included both official and unofficial translations, as newer generations and *Fakémon* (fan-created *Pokémon*) often lack official localised names. Most

of the names in the corpus are portmanteau words – i.e., words formed by merging the sounds and meanings of two different words.

For this year's task, we doubled the size of the corpus and added explanatory descriptions in English for each instance of wordplay, sourced variously from Wikipedia, the Web, and Master's students in translation.

For training purposes, we released 353 onomastic wordplay instances in English with corresponding French translations and descriptions, all drawn from *Asterix* and *Harry Potter*. These sources are well known and well documented on Web sources such as Wikipedia. For testing purposes, we compiled our own dataset of instances manually translated by trained professionals, as well as instances of official translations. We used 2,333 instances of onomastic wordplay in English with corresponding French translations as our test set.

3.2. Format

Input. We provide the training data as JSON 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 onomastic 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
- **description**: short contexts for the names and objects, which are often necessary to recognise, understand, and translate the wordplay
- **fr**: translation of the onomastic wordplay into French

For example:

```
[
  {
    "id":"en_1",
    "en":"Asterix",
    "description":"Asterix is the small but clever hero of the Asterix comic series.
        Known for his sharp wit and courage, he outsmarts the Roman invaders with the
        help of a magical potion that grants him superhuman strength. Alongside his
        loyal friend Obelix, Asterix defends his village and embodies bravery and
        cleverness.",
    "fr":"Astérix"
  }
]
```

The test data format is identical to that of the training data, except that the field for the target text is omitted.

Output. Participants were asked to submit to Codabanch a ZIP archive containing a file named `prediction.json` in the root directory. This JSON-formatted file was to contain the following fields:

- **run_id**: Run ID starting with `<team_id>_<task_id>_<method_used>` – e.g., `UBO_task_3_BLOOM`
- **manual**: 0 if the run is automatic, or 1 if manual
- **id_en**: a unique identifier from the input file
- **en**: the text of the instance of source onomastic wordplay in English.
- **fr**: translation of the onomastic wordplay into French

Example:

```
[
  {
    "run_id":"team1_task_3_DeepL",
    "manual":0,
    "id_en":"en_1",
    "en":"Asterix",
    "fr":"Astérix"
  }
]
```


4. Evaluation

For the wordplay translation, there do not yet exist any universally accepted metrics of translation quality [21, 22]. Machine translation is traditionally measured with the BLEU (Bilingual Evaluation Understudy) metric, which calculates vocabulary overlap between the candidate translation and a reference translation [23]. However, this metric is clearly inappropriate for single-term wordplay translation evaluation, as overlap measures operate only on larger text spans and not on individual words, the morphological analysis of which can be crucial for neologisms [21, 22]. We therefore evaluated participants' translations by automatically checking them for case-insensitive exact matches against the manual reference translations; we report these scores under the label "automatic".

We hypothesised that the majority of proper nouns would not be translated automatically, so we also checked the target translations for identity with the source texts, and report these scores under the label "identical".

Finally, we performed a manual evaluation of 1,737 translations of 203 distinct source wordplay instances sampled from the participants' runs. The annotation was carried out by Master's students in translation, who evaluated whether the generated translations were valid alternatives – i.e., whether they preserved the wordplay and conveyed a meaningful name for the character or object in context. Descriptions and reference translations were also provided to support this task. Although we tried to remove translations matching the references and the English sources to reduce the annotators' working load, some of them were maintained due to the slight format differences. We added reference translations to calculate the percentage of successful translations in runs resulting in 1,833 distinct lower-cased stripped translations. These manual evaluation scores are reported under the label "manual".

5. Participants' approaches

Four teams participated in this task for a total number of 20 runs:

mariapazr20 [13] This team used chain-of-thought prompting techniques with several large language models, including additional constraints identified from recurring translation patterns for each literary work of the provided corpus (such as favouring puns in a given semantic field over meaning preservation for instance).

arampageos [12] This participant started by manually or semi-automatically classifying the names in the training data into four categories (alliteration, wordplay, realistic names and unclassified). They then used the same strategy as for Task 2, with a two-stage approach where a record of manually defined translations preceded using large language models. They applied different machine translation models (e.g., MarianMT, Helsinki-NLP-opus, facebook-nllb, T5).

sarath_kumar [15] This participant prepared a dataset containing named entity recognition annotations and used it to source translations from the T5-base model. They used a beam search to prioritise

phonetic matches between source and target names, and then ranked the translations according to their creativity, phonetic fidelity, and cultural relevance.

pjmathematician [14] This team used zero-shot prompting with Qwen models. The prompt consists of about 50 lines and includes guidance on how to translate wordplay (when to not translate, when to use a literal translation, or suggesting creative constraints such as considering characters' traits or relying on the story universe vocabulary).

All participants who submitted runs also submitted system description papers to the Working Notes volume [24]. 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.

6. Results

6.1. Test data

Table 6.1 reports the percentage of matching translations for each run, according to the aforementioned manual and automatic metrics. For context, it also reports the percentage of instances in each run where the translated French wordplay is identical to the English original; this figure is 100% for the run labelled "copy", is a naïve baseline that "translates" by copying input to output verbatim.

The best run according to both manual evaluation (62.56%) and exact match to the references (39%) is VerbaNex_gpt4o [13]. Following this, with about half the exact-match score, are two runs by the team pjmathematician [14]. However, the differences are much less stark according to the manual evaluation, which allowed alternative translations (62.56% vs. 46,31%). In both teams' top-scored runs, VerbaNex_gpt4o and pjmathematician_Q332, we observe that 23% of generations were considered to be successful alternative translations. The top-scored VerbaNex_gpt4o has only 8.53% of translations identical to the English source wordplay.

Translation analysis. About 12% of reference translations are identical to the English source wordplay. The manual evaluation shows only 2.55% of untranslated wordplay instances as appropriate translations, as we tried to remove translations matching the references and the English sources in order to reduce the annotators' working load. These 2.55% correspond to some translations identical to the source that were not filtered out due to the minor differences in formatting or typography. The identity baseline remains a strong one, outperforming more than half of submitted runs in terms of matching to the references. Half the runs have more than 40% French translations identical to English while 30% keep half the onomastic wordplay instances untranslated. This proportion is much higher than in the references but aligns with traditional approaches to translating named entities that omit wordplay [9, 10], which often fail to preserve the intended humorous or pragmatic meaning for the target audience.

Among 1,737 manually evaluated translations, 172 (10%) were considered successful ones. Among these, 17 were nearly identical to the reference translations, with the only differences manifesting in diacritics, capitalisation, and/or punctuation – for example, "Oreilles de Soie" (run) vs. "Oreilles-De-Soie" (reference) for "Ears of Silk" (source). Less than 10% of manually evaluated translations (155 instances) were genuinely alternative translations, suggesting that translating onomastic wordplay remains a challenge despite the impressive capacities of LLMs. Among recurrent errors, 226 generations were identical to the English source. In 102 cases, we found the suffix "-ix" as in Celtic names, which might be a result of overfitting on the training set containing the names from the *Asterix* comics. In 11 cases, the translations lack the character's surname. Twenty-nine generations were blank or consisted only of punctuation (e.g., "???"), and in 13 cases we found spurious overgeneration such as "l'aide de" or seemingly random translations such as "l'intention des autorités fédérales, il" for "Chimchar". There were 226 translations containing extraneous articles (*le, l', la, les*) as in "Le Munchlax" for "Munchlax"

Table 2
Results for Task 3 on the test data, showing the percentage of matching translation instances according to the automatic and manual evaluations, as well as the percentage of translation instances identical to the source text

run ID	automatic	manual	identical
VerbaNex_gpt4o	39.05	62.56	8.53
pjmathematician_Q332	22.85	46.31	21.82
pjmathematician_Q314	21.13	39.60	33.48
duth_Helsinki	14.83	18.88	77.67
duth_xanthi_Helsinki-NLP-opus-mt-tc-big-en-fr	14.66	18.88	77.45
Cryptix_flanT5	14.49	13.43	38.15
duth_Helsinki	11.83	2.55	100.00
duth_xanthi_facebook-nllb-200-distilled-600M	10.72	16.75	41.83
duth_xanthi_facebook-nllb-200-1.3B	10.72	16.75	41.83
duth_xanthi_MarianMT_BLOOM	10.42	13.86	45.95
duth_xanthi_MarianMT_BLOOM	10.29	13.86	45.78
duth_xanthi_Helsinki-NLP-opus-mt-en-fr	10.29	13.86	45.78
duth_xanthi_t5-base	8.57	7.03	50.32
duth_xanthi_t5-small	8.53	6.00	58.04
duth_xanthi_facebook-m2m100_1.2B	4.71	9.50	19.12
duth_xanthi_facebook-m2m100_418M	4.37	4.00	20.15
team1_gemma2b_v2	4.20	2.99	27.69
duth_hybrid_v1	0.04	1.47	0.21
duth_xanthi_MarianMT_LLM_Prompting	0.00	0.00	0.00
Skommarkhos_Lucie-7B-Instruct_SFT_Q8B_LoRA	0.00	0.00	0.00
copy	11.83	2.55	100.00

or “Le Shinx” for “Shinx”. The Pokémon name “Pidove” was inexplicably translated as “pédophile” in one run.

6.2. Training data

As for other tasks, runs were submitted for both the training and test datasets in order to analyse potential overfitting and related effects. Tables 6.2 report the Task 3 results on the training data, showing the percentage of translations matching the references and the percentage of translation instances identical to the source text. Unlike the test data, the training data was not manually evaluated due to cost constraints.

Of the 21 runs, 14 (33%) submitted by teams duth [12] and pjmathematician [14] achieved nearly identical scores by both exactly matching the reference translations (56% of successful translations) and retaining the English source text as the translation (3.4%). These runs implement very different models and show varied performance on the test data. Submitted translations on the training data show a much higher rate of exact matches with the reference translations compared to the test set, while the number of untranslated names is markedly lower. The best-scoring run on the training data, pjmathematician_Q332 [14], shows a drop in exact matches with the reference translations from 56% to 23%, while the proportion of untranslated names increases from 3.4% to 22% on the test data. The training data was sourced from *Asterix* and *Harry Potter*, which have well-known official translations; for the test data, we created completely new data. The higher rate of exact matches and lower number of untranslated names in the training data may be explained by the potential inclusion of this data in the training sets of AI models.

Surprisingly, VerbaNex_gpt4o does not follow this trend; it had only 12% of exact matches on the training data while it achieved 39% on the test. The percentage of untranslated names for VerbaNex_gpt4o does not differ a lot between the training and test data. Note that according to manual evaluation, 63% of translations were successful. This might be explained by more creative alternative translations.

Table 3
Results for Task 3 on the training data, showing the percentage of translations matching the references and the percentage of translation instances identical to the source text

run ID	automatic	identical
pjmathematician_Q332	55.81	3.40
pjmathematician_Q314	55.81	3.40
duth_xanthi_facebook-m2m100_418M	55.52	3.40
duth_xanthi_Helsinki-NLP-opus-mt-en-fr	55.52	3.40
duth_hybrid_v1	55.52	3.40
duth_xanthi_t5-base	55.52	3.40
duth_xanthi_MarianMT_BLOOM	55.52	3.40
duth_xanthi_Helsinki-NLP-opus-mt-tc-big-en-fr	55.52	3.40
duth_xanthi_MarianMT_LLM_Prompting	55.52	3.40
duth_xanthi_facebook-nllb-200-distilled-600M	55.52	3.40
duth_xanthi_facebook-m2m100_1.2B	55.52	3.40
duth_xanthi_t5-small	55.52	3.40
duth_xanthi_facebook-nllb-200-1.3B	55.52	3.40
duth_Helsinki	55.52	3.40
VerbaNex_gpt4o	11.61	10.48
Cryptix_flanT5	8.50	29.46
duth_Helsinki	5.38	52.69
Skommarkhos_CroissantLLMChat-v0.1_SFT_Q8B_LoRA	4.53	26.35
Skommarkhos_Lucie-7B-Instruct_SFT_Q8B_LoRA	3.97	5.10
team1_gemma2b_v2	1.42	17.28
copy	0.00	0.00

Further analysis is needed to prove this hypothesis.

7. Conclusion

For Task 3, we constructed a parallel corpus of English and French onomastic wordplay collected from video games, slogans, literature, and other sources with translations and context descriptions, which are often necessary to recognise, understand, and translate the wordplay. We released 353 training instances of English onomastic wordplay with corresponding translations into French and descriptions. For testing, we assembled 2,333 English wordplay instances paired with professional or official French translations and descriptions providing necessary context. Participants' submissions were evaluated both automatically by exact matching and manually on 1,737 translations, with some near-identical outputs retained due to minor formatting or typographical differences, resulting in 1,833 distinct normalised translations for analysis.

Four teams submitted a total number of 20 runs to Codabench. Fourteen runs (33%) from teams duth [12] and pjmathematician [14] achieved nearly identical scores, with high exact matches (56%) and few untranslated names (3.4%) on the training data. However, the best run, pjmathematician_Q332, dropped to 23% exact matches and 22% untranslated names on the test set. This likely reflects differences in data composition: the training set drew on well-known, officially translated sources like *Asterix* and *Harry Potter*, whereas the test set included partially new data, reducing potential overlap with AI training data.

Surprisingly, VerbaNex_gpt4o achieved only 12% exact matches on the training data but 39% on the test, with a stable rate of untranslated names and 63% successful translations by manual evaluation, possibly due to more creative alternative translations—a hypothesis requiring further analysis.

Half the runs left more than 40% of the instances untranslated. This proportion is much higher than in the reference corpus, which has 12% of translations identical to the source, but aligns with traditional approaches to translating named entities that omit wordplay [9, 10]. This traditional approach may

nonetheless fail to preserve the intended humorous or pragmatic meaning for the target audience.

Despite notable advances, translating onomastic wordplay remains highly challenging, with fewer than 10% of manually evaluated translations judged as genuine alternatives. While VerbaNex_gpt4o achieved the highest performance overall, a significant portion of the outputs of runs still relied on identity to the English source or near-verbatim adaptations, illustrating the limitations of current models. Recurrent errors – such as untranslated names, overfitting to training data, omission of surnames, and occasional nonsensical generations – highlight that further progress is needed to produce creative, culturally adapted translations at scale.

In the future, we plan to perform more detailed analysis of the alternative generated translations and compare human and machine strategies for neologism creation as well.

For more information about the JOKER lab this year, please refer to the overview paper [1] and to the task description papers for Task 1: Humour-aware Information Retrieval [4] and Task 2: Translation of Puns from English to French [5]. Visit the JOKER website at https://joker-project.com for any other information related to the track.

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, 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.

[3] L. Ermakova, T. Miller, J. Boccou, A. Digue, A. Damoy, P. Campen, Overview of the CLEF 2022 JOKER task 2: Translate wordplay in named entities, 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. 1666–1680. URL: https://ceur-ws.org/Vol-3180/paper-127.pdf.

[4] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the CLEF 2025 JOKER Task 1: Humour-aware Information Retrieval, in: [24], 2025.
[5] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 JOKER Task 2: Wordplay Translation from English into French, in: [24], 2025.
[6] J. J. O'Hara, True names: Vergil and the Alexandrian tradition of etymological wordplay, University of Michigan Press, 2017.
[7] R. C. Beshere, "What's in a name?": Theorizing an etymological dictionary of Shakespearean characters, The University of North Carolina at Greensboro, 2009.
[8] L. Manini, Meaningful literary names: Their forms and functions, and their translation, The Translator 2 (1996) 161–178.
[9] N. Chen, R. E. Banchs, M. Zhang, X. Duan, H. Li, Report of NEWS 2018 named entity transliteration shared task, in: Proceedings of the Seventh Named Entities Workshop, Association for Computational Linguistics, 2018, pp. 55–73. doi:10.18653/v1/W18-2409.
[10] N. Chen, X. Duan, M. Zhang, R. E. Banchs, H. Li, NEWS 2018 whitepaper, in: Proceedings of the Seventh Named Entities Workshop, Association for Computational Linguistics, 2018, pp. 47–54. doi:10.18653/v1/W18-2408.
[11] 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.
[12] G. Arampatzis, A. Arampatzis, DUTH at CLEF JOKER 2025 Tasks 2 and 3: Translating Puns and Proper Names with Neural Approaches, in: [24], 2025.
[13] M. P. R. Atencio, J. D. Jimenez, D. Gómez, J. E. Serrano, E. Puertas, VerbaNexAI at CLEF 2025 JOKER Task 3: Multi-Model LLM Approach for Onomastic Wordplay Translation, in: [24], 2025.
[14] 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: [24], 2025.
[15] 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: [24], 2025.
[16] P. Zabalbeascoa, S. Attardo, Humour translation theories and strategies, in: L. Kostopoulou, V. Misiou (Eds.), Transmedial perspectives on humour and translation: from page to screen to stage, Advances in Translation and Interpreting Studies, Routledge, New York/London, 2024.
[17] J. Vandaele, Translating literary humour: Aspects of detection and analogy making 1, in: L. Kostopoulou, V. Misiou (Eds.), Transmedial perspectives on humour and translation: from page to screen to stage, Advances in Translation and Interpreting Studies, Routledge, New York/London, 2024.
[18] V. Chan, W. K.-W. Tang, GPT and translation: A systematic review, in: 2024 International Symposium on Educational Technology (ISET), IEEE, Macau, Macao, 2024, pp. 59–63. doi:10.1109/ISET61814.2024.00021.
[19] H. Abu-Rayyash, AI meets comedy: Viewers' reactions to GPT-4 generated humor translation, Ampersand 12 (2024) 100162. doi:10.1016/j.amper.2023.100162.
[20] R. Hamilton, Artificially funny: collaborative play at the intersection of AI, literature and humour, in: W. Slocombe, G. Liveley (Eds.), The Routledge handbook of AI and literature, Routledge literature handbooks, Routledge, New York, 2025.
[21] L. Ermakova, T. Miller, O. Puchalski, F. Regattin, É. Mathurin, S. Araújo, A.-G. Bosser, C. Borg, M. Bokiniec, G. L. Corre, B. Jeanjean, R. Hannachi, Ġ. Mallia, G. Matas, M. Saki, CLEF Workshop JOKER: Automatic Wordplay and Humour Translation, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, volume 13186 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022, pp. 355–363. doi:10.1007/978-3-030-99739-7_45.
[22] 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.

[23] 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. doi:10.3115/1073083.1073135.

[24] 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.