University of Amsterdam at the CLEF 2025 JOKER Track

Alecsandru Kreefft-Libiu, Finley Helms, Cem Selçuk, Jan Bakker and Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper reports on the University of Amsterdam's participation in the CLEF 2025 JOKER track. Our overall goal is to investigate the non-literal use of language, such as in humor and wordplay, that remains challenging for current information retrieval and natural language processing technologies. Our specific focus is on a post hoc approach where we exploit wordplay or humorous text detection as a means to filter out humorous translations or search results. Our main findings are the following. First, we successfully developed effective humor detection classifiers for both English and French. Second, for humor-aware information retrieval, we could increase retrieval effectiveness by filtering for humorous content in search results ranked solely on topical relevance. Third, for wordplay translation, we manage to generate multiple translation candidates and select the one with the highest pun score based on the detector. This approach performs well, with limited gain on the automatic evaluation measures, but qualitative analysis confirms that this is an encouraging strategy.

Keywords

Information Storage and Retrieval, Natural Language Processing, Wordplay Translation, Humor Retrieval, Funny Names Translation

1. Introduction

The CLEF 2025 JOKER track investigates possible solutions to the challenges of automated analysis and processing of humor. The JOKER track series aims to advance the development of interpretation, generation, and translation of wordplay by bringing together computer scientists, linguists, and translators. The CLEF 2025 JOKER Track builds upon the findings from CLEF 2024 JOKER Track [1], in particular on humor-aware search [2], and wordplay translation [3]. We conduct an extensive analysis of the track: Task 1 on *Humor-aware Information Retrieval*; Task 2 on *Wordplay Translation*; and Task 3 on *Onomastic Wordplay Translation*. For details on the exact track setup in 2025, we refer to the Track Overview paper CLEF 2025 LNCS proceedings [4], as well as the detailed task overviews in the CEUR proceedings [5, 6, 7].

Our main aim is to investigate a post hoc approach that uses wordplay or humorous text detection to filter out humorous translations or search results. For example, as discussed in detail below, we can utilize a pun detector to filter wordplay from the results of a standard search engine, focusing solely on topical relevance. Similarly, we can also have standard machine translation systems generate different translation candidates and detect which of them preserve the wordplay. The key element in our humor-aware approach is to utilize effective classifiers for puns and wordplay.

The rest of this paper is structured as follows. Next, in Section 2 we discuss our experimental setup and the specific runs submitted. Section 3 discusses the results of our runs. We end in Section 4 by discussing our results and outlining the lessons learned.

2. Experimental Setup

In this section, we will detail our approach for the three CLEF 2025 JOKER track tasks.

For details of the exact task setup and results, we refer the reader to the detailed overview of the track in [4] and [1]. The basic ingredients of the track are:

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

j.bakker@uva.nl (J. Bakker); kamps@uva.nl (J. Kamps)

0009-0002-9085-8491 (J. Bakker); 0000-0002-6614-0087 (J. Kamps)

Table 1
CLEF 2025 JOKER Track Submissions

	Run	Description
1	UAms_en_bm25	BM25 baseline (Anserini, stemming)
1	UAms_en_rm3	RM3 baseline (Anserini, stemming)
1	UAms_en_bm25_CE1K	BM25 + Crossencoder top 1,000
1	UAms_en_rm3_CE1K	BM25/RM3 + Crossencoder top 1,000
1	UAms_RM3	Okapi BM25/RM3
1	UAms_RM3RoBERTa	BM25/RM3 + Filter on RoBERTa Pun classifier (keeps 90%)
1	UAms_RM3RoBERTa_drop60	BM25/RM3 + Filter on RoBERTa Pun classifier (keeps 40%)
1	UAms_pt_bm25	BM25 baseline (Anserini, stemming)
1	UAms_pt_rm3	RM3 baseline (Anserini, stemming)
1	UAms_pt_bm25_CE1K	BM25 + Crossencoder top 1,000
1	UAms_pt_rm3_CE1K	BM25/RM3 + Crossencoder top 1,000
2	UvA_finetunedmBARTcc25	mBART-large-cc25 Finetuned
2	UvA_mBARTcc25&finetunedroBERTa	mBART-large-cc25 + humour detector roBERTa-large filter
2	UvA_finetunedT5-base	T5-base Finetuned
2	UvA_T5-base&finetunedroBERTa	T5-base + humour detector roBERTa-large filter
2	UvA_finetunedNLLB-1.3B	NLLB-200-1.3B Finetuned
2	UvA_finetunedNLLB-1.3B&finetunedroBERTa	NLLB-200-1.3B + humour detector roBERTa-large filter
2	UvA_finetunedMarianMT	MarianMT Finetuned
2	UvA_finetunedMarianMT&finetunedroBERTa	MarianMT + humour detector roBERTa-large

Corpus For Task 1, there is a large corpus of 77,658 documents (usually a single sentence each) for the retrieval task in English. There is an additional Portuguese corpus of 45,126 documents (including a small fraction of AI-generated humorous texts).

Train Data For Task 1, there are 12 English train queries with relevance judgments (between 4 and 364 relevant per query). There are 29 Portuguese train queries with relevance judgments (between 2 and 40 relevant per query).

For Task 2, there are 1,405 English wordplays, with a total of 5,838 professional human translations in French.

For Task 3, there are 353 English onomastic wordplays, or informally "funny" names, with professional human translations in French.

Test Data For Task 1, there are 219 English test queries. These include the 12 train queries, resulting in a total of 207 unseen queries on which the test evaluation is based. For these unseen queries, there are between 1 and 233 relevant documents. There are also 98 Portuguese test queries, not including any of the train queries, with between 1 and 79 relevant documents.

For Task 2, there are 1,682 English wordplays, with 2,615 reference translations into French by professional translators.

For Task 3, there are 2,333 English onomastic wordplays, with French reference translations made by professional translators

We created runs for two of the tasks of the 2025 track, which we will discuss in order.

2.1. Task 1: Humor-aware Information Retrieval

This task asks to retrieve short humorous texts for a query. We submitted twelve runs in total, shown in Table 1.

Baseline Rankers We first submitted four baseline runs focusing on regular information retrieval effectiveness. Two are vanilla baseline runs on an Anserini index, using either BM25 or BM25+RM3 with default settings [8].[1] The other two runs are neural cross-encoder rerankings of these runs, based on zero-shot application of an MSMARCO trained ranker, reranking the top 1,000 of either the BM25 or the BM25+RM3 baseline run.[2] We submitted four runs for both the English and the Portuguese data. To understand the effectiveness of standard retrieval systems optimized for topical relevance, our submissions used default, non-optimized settings and privileged recall over precision.

RoBERTa All three submitted runs were based on Okapi BM25 as the retrieval model, combined with RM3 relevance feedback. The first run, `UAms_RM3`, applied no filtering and served as a BM25+RM3 baseline. The second run, `UAms_RM3RoBERTa`, added a filtering step using a pre-trained RoBERTa-based pun classifier.[3] Documents with a predicted pun probability above a tuned threshold were retained, resulting in approximately 90% of documents being kept. The third run, `UAms_RM3RoBERTa_drop60`, used the same classifier, but instead of threshold tuning, it applied a fixed-ratio filter that retained the top 40% of documents with the highest predicted pun probabilities.

2.2. Task 2: Wordplay Translation

This task asks to translate english punning jokes into french. We submitted eight runs, as shown in Table 1.

MarianMT MarianMT is a "sequence-to-sequence" (Seq2Seq) model based on the Marian framework. Marian, first introduced in 2017, is written entirely in C++, which supports faster training and translation. MarianMT provides pre-trained models, which are smaller than most other translation models, about 298 MB on disk, compared to other transformer-based translation models that exceed 1 GB. The size of MarianMT makes the model useful for fine-tuning on custom datasets for specific tasks. For the translation task, the MarianMTModel and MarianTokenizer were loaded from the transformers library, using the model checkpoint "Helsinki-NLP/opus-mt-en-fr". Before training, the data was preprocessed by merging the input en qrels files on "id_en", to create a single csv file. The columns were renamed to "English" and "French", no prefix was needed. The preprocessed data was divided with train_test_split into a 90/10 split of respectively training and test data. After which the training data was further divided into training and validation sets using a 80/20 split.

T5-base As stated in previous paragraphs, a T5 model can be used for different NLP tasks. It is suitable for machine translation due to its ability to understand natural language and generate contextually relevant information. The 'T5ForConditionalGeneration' and the model name 't5-base' were used to load the T5-base model for English to French translation. The preprocessing of the data was done similarly to the preprocessing for the MarianMT model. The split of the test, validation and train set was also done in the same manner. The 'T5Tokenizer' was used to tokenize the data before training. Training was done with the same number of epochs, batch size and evaluation metric as used for MarianMT.

NLLB-200-1.3B The NLLB-200-1.3B model (No Language Left Behind), developed by Meta AI, is a state-of-the-art multilingual translation model supporting 200 languages. It uses an optimized Transformer-based encoder-decoder architecture to produce high-quality translations, especially for low-resource languages. For the translation task, the model "facebook/nllb-200-1.3B" was loaded from the transformers library, along with the corresponding NllbTokenizer. Since NLLB requires explicit language codes, the input sentence was prefixed with the target language ID, in this case `fra` for French.

[1] https://github.com/castorini/pyserini
[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
[3] https://huggingface.co/frostymelonade/roberta-small-pun-detector-v2

mBART-large-cc25 The mBART (Multilingual BART) model, introduced by Facebook AI, is an encoder-decoder model that combines multilingual pretraining and fine-tuning. The mbart-large-cc25 variant is pretrained on 25 languages using a denoising autoencoding objective, allowing it to generalize well for translation tasks. For this translation task, the model facebook/mbart-large-cc25 was used from the transformers library, along with the MBartTokenizer. Like NLLB, mBART requires explicit language codes. For English-to-French translation, en_XX was used as the source language code and fr_XX as the target language code. These tokens were appended to the input during encoding and decoding respectively.

2.2.1. Humour Detection & Translation

Building on the foundation laid by the UvA's participation in the 2024 CLEF JOKER Track [9], while literal translations can often be handled effectively by LLMs with sufficient training, the translation of puns and other forms of wordplay presents a considerably greater challenge due to their inherently ambiguous and context-dependent nature. To explore new possibilities, we employed the same RoBERTa-based model as described in the previous section and fine-tuned it on the binary pun detection task presented in the 2023 CLEF JOKER Track [10]. This task focuses on distinguishing sentences that contain puns from those that do not, providing a useful benchmark for evaluating wordplay sensitivity in NLP systems. For detailed information regarding the dataset and task formulation, we refer the reader to the official workshop notes of the Pun Detection Task [10].

Subsequently, we integrate this pun detector with each of the MT models under investigation. This setup allows us to systematically assess how pun-aware pre-processing could influence the performance and behavior of MT systems. Following the methodology of the CLEF 2024 track approach [9], we generate three candidate translations for each pun-containing sentence using beam search in MarianMT. Each of these translations is then evaluated using our fine-tuned pun classifier, which assigns both binary class labels ("pun" or "non-pun") and associated class probabilities. The candidate translation with the highest pun probability score was selected as the final output.

2.3. Task 3: Onomastic Wordplay Translation

This task asks to translate funny names from english into french. We made no official submissions for this task other than trial runs with the Task 2 approach applied to the names and the descriptions of the onomastic wordplay.

3. Experimental Results

In this section, we will present the results of our experiments in self-contained subsections following the CLEF 2025 JOKER Track tasks.

3.1. Task 1: Humor-aware Information Retrieval

We discuss our results for Task 1, asking to retrieve short humorous texts for a query.

Table 2 shows the performance of the Task 1 submissions on the test data. None of our approaches was trained or informed by the train data, nor was pooling used to locate relevant documents (the corpus's recall base is known to be complete). As a consequence, the results of the train and test data are comparable. First, we observe that standard lexical ranking approaches perform reasonably, albeit not very well, and do not yield a significant performance gain when using blind feedback. In particular, for Portuguese, the scores are low, possibly also due to the synthetic nature of the corpus construction. We also see that the neural rankers may attract topically relevant content, but do not attract non-humorous content. As a result, zero-shot rankers lead to a decrease in performance due to the large fraction of non-relevant documents. Second, the filter based on a pun classifier is effective, leading to a notable

Table 2
Evaluation of JOKER Task 1 (test data).

Run	MAP	GMAP	P@R	MRR	Precision				NDCG
					5	10	100	1,000	5
UAms_en_bm25	11.91	5.64	12.23	26.28	12.95	12.71	5.94	0.91	14.00
UAms_en_rm3	10.76	5.35	10.56	25.03	11.88	12.22	6.14	1.03	12.37
UAms_en_bm25_CE1K	4.88	2.60	2.47	5.68	0.48	2.75	4.30	0.91	0.38
UAms_en_rm3_CE1K	4.78	2.67	2.32	5.48	0.39	2.51	4.32	1.03	0.27
UAms_RM3	15.02	7.22	19.53	40.87	24.35	20.00	6.22	0.90	25.66
UAms_RM3RoBERTa	14.94	7.16	19.56	42.47	25.51	19.61	6.07	0.87	26.76
UAms_RM3RoBERTa_drop60	16.72	7.04	23.05	54.46	30.82	23.09	5.98	0.70	1.52
UAms_pt_bm25	7.89	0.19	5.96	9.83	5.22	6.09	3.03	0.33	5.13
UAms_pt_rm3	6.54	0.25	5.91	9.51	4.64	5.65	2.78	0.38	4.47
UAms_pt_bm25_CE1K	3.84	0.12	1.99	4.77	1.16	3.04	2.35	0.33	0.91
UAms_pt_rm3_CE1K	4.16	0.19	2.47	5.20	1.45	3.19	2.41	0.38	1.34

Table 3
CLEF 2025 JOKER Task 2: Test results

Run	Location	BLEU	Precisions				BERTScore		
			1	2	3	4	P	R	F1
UvA_finetunedmBARTcc25	3.80	16.55	39.64	19.49	12.22	7.95	79.48	79.79	79.59
UvA_mBARTcc25&finetunedroBERTa	5.29	18.20	40.86	21.09	13.74	9.26	80.07	80.00	80.00
UvA_finetunedT5-base	6.30	36.77	60.29	40.58	30.94	24.15	86.69	86.24	86.44
UvA_T5-base&finetunedroBERTa	6.36	36.14	59.75	39.92	30.30	23.61	86.42	86.12	86.24
UvA_finetunedNLLB-1.3B	6.36	42.55	64.74	46.26	36.70	29.83	87.85	87.04	87.42
UvA_finetunedNLLB-1.3B&finetunedroBERTa	6.60	41.80	63.86	45.49	36.01	29.17	87.55	86.96	87.23
UvA_finetunedMarianMT	6.78	41.19	63.37	44.74	35.31	28.76	87.72	86.92	87.24
UvA_finetunedMarianMT&finetunedroBERTa	6.78	40.85	62.90	44.38	35.03	28.49	87.50	86.78	87.11

increase in both precision and recall (MRR, NDCG, and MAP). We observe that the more aggressive filter, which retains 40% of results, outperforms the moderate filter, which retains 90% of results.

3.2. Task 2: Wordplay Translation

We continue with Task 2, asking to translate english punning jokes into french. We submitted eight runs, using four stand-alone finetuned models and four stand-alone models combined with a pun-detector.

Official Evaluation Results Table 3 shows the results of the CLEF 2025 JOKER track's Task 2 on the test data. We make a number of observations.

First, the general translation quality varies across systems, with BLEU scores ranging from 16.55% to 42.55% and BERTScore F1 ranging from 79.59% to 87.42%. These results demonstrate that some models are capable of generating fluent and accurate translations, although preserving the pun remains a challenge. Second, we observe that the inclusion of the pun detector has a small effect on BERTScore F1, slightly improving results in mBARTcc25. Its combination with the fine-tuned RoBERTa achieves a BERTScore F1 of 80.00%, compared to 79.59% without the detector. This trend, however, does not hold for all models, and performance seems to decrease slightly on these metrics. Third, examining the location of the punword and whether the output exactly contains this punword from the reference translation, we observe an increasing trend, albeit with a small gain and a draw for the best-performing MarianMT model.

Table 4
CLEF 2025 JOKER Task 2: Example (id_en: en_1007 in train, and en_1 in test)

Run	Text
Source	Save the whales, spouted Tom.
Reference(s)	“Il faut sauver les baleines,” jeta Tom avant de se tasser. “Il faut sauver les baleines,” interjeta Tom. Moi je sauve les baleines, Tom s’en venta. Louis évent-a le projet de sauvetage des baleines. “Sauvez les baleines,” proclama Tom à tout évent. “Sauvez les baleines, cracha Toto, Cétacé!”
UvA_finetunedmBARTcc25	”Les baleines, ..., les baleines, les baleines.”
UvA_mBARTcc25&finetunedroBERTa	”Les baleines, les baleines, s’exclama Tom”
UvA_finetunedT5-base	”Sauver les baleines, dit Tom.”
UvA_T5-base&finetunedroBERTa	”Sauver les baleines, dit Tom.”
UvA_finetunedNLLB-1.3B	”Šauvez les baleines, s’exclama Tom.”
UvA_finetunedNLLB-1.3B&finetunedroBERTa	”Šauvez les baleines, s’exclama Toto.”
UvA_finetuneMarianMT	”Sauver les baleines, a filé Tom.”
UvA_finetuneMarianMT&finetunedroBERTa	”Sauver les baleines, a lancé Tom.”

Automatic evaluation metrics assess the entire translated sentence as a whole. They are a necessary but not sufficient condition for successful pun translation. The ground truth consists of professional translations that preserve the wordplay across languages. Therefore, while the results are acceptable, they also underline the importance of further qualitative analysis of the translations.

Qualitative Evaluation Table 4 shows an example from the train data set. The top half of the table shows the English pun and the six French translations made by professional translators. The bottom half of the table shows the translations generated by our systems. We make a number of observations:

First, most model outputs match the style and wordplay of professional references quite well, especially when the pun detector is applied. Every model takes something from the original. It is the way the verb ``Sauve`` is conjugated or the way Tom expresses himself that is changed. Notably, the `UAms_Task2_NLLB-200-1.3B_roBERTa-large_filter` has chosen for *Toto* as the correct translation of Tom, even though it only comes forth once in the references. Toto jokes are a common type of joke in French,[4] similar to Tom (also Tom Swifty) jokes in English.[5]

Second, other outputs (e.g., from T5-base) are fluent and grammatically correct, but they fail to preserve the pun. These outputs often resemble literal translations that miss the wordplay entirely, which can occur when only the top-ranked translation is used.

Our analysis revealed both the quality of current machine translation and the complexity of preserving the wordplay in a literally correct translation. We also observed that the models are able to generate creative translations preserving the wordplay, but that the most likely translation or the first one generated by the model may not be a pun. This observation supports our general idea to generate multiple translations with the model and use an effective pun detector to choose one of these translations in case it is likely to preserve the wordplay.

4. Discussion and Conclusions

This paper detailed the University of Amsterdam’s participation in the CLEF 2025 JOKER track. We conducted a range of experiments for each of the three tasks of the track. Our primary objective was to develop a post hoc approach that leverages wordplay or humorous text detection to filter out humorous translations or search results. Our main findings are the following. First, we successfully

[4]https://fr.wikipedia.org/wiki/Blague_de_Toto
[5]https://en.wikipedia.org/wiki/Tom_Swifty

developed effective humor detection classifiers for both English and French. Second, for humor-aware information retrieval, we could increase retrieval effectiveness by filtering for humorous content in search results ranked solely on topical relevance. Third, for wordplay translation, we generated multiple translation candidates and selected the one with the highest pun score, as determined by the detector. This approach performed well, with limited gain on the automatic evaluation measures, and qualitative analysis confirmed that this is an encouraging strategy.

Our qualitative analysis also demonstrated the importance and potential of models that capture such deep cultural references, such as the Wellerism of Tom Swifty jokes that match French *Blague de Toto* wordplay. At the same time, it also highlights the complexity of evaluating output against these references: this particular cultural reference hinges on only a single word in one of the reference translations. The importance of generating high-quality professional data sets from human translators is paramount.

Acknowledgments

This research was conducted as part of the final research projects of the Bachelor in Artificial Intelligence at the University of Amsterdam, We thank the coordinator, Dr. Sander van Splunter, for his support and flexibility to work around the CLEF deadlines. We also thank the track and task organizers for their amazing service and effort in making realistic benchmarks available for analyzing and processing humorous text.

Jan Bakker is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is supported by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

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

References

[1] L. Ermakova, A. 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, G. M. D. Nunzio, L. Soulier, P. Galuscáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 165–182. URL: https://doi.org/10.1007/978-3-031-71908-0_8. doi:10.1007/978-3-031-71908-0_8.

[2] L. Ermakova, A. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1775–1785. URL: https://ceur-ws.org/Vol-3740/paper-165.pdf.

[3] L. Ermakova, A. 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. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1800–1810. URL: https://ceur-ws.org/Vol-3740/paper-167.pdf.

[4] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 Joker track: Humour in the machine, in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco 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.

[5] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 Joker Task 1: Humous-Aware Information Retrieval , in: [11], 2025.

[6] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 Joker Task 2: Wordplay Translation, in: [11], 2025.

[7] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 Joker Task 3: Onomastic Wordplay Translation, in: [11], 2025.

[8] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, R. F. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2356–2362. URL: https://doi.org/10.1145/3404835.3463238. doi:10.1145/3404835.3463238.

[9] E. Schuurman, M. Cazemier, L. Buijs, J. Kamps, University of amsterdam at the CLEF 2024 joker track, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1909–1922. URL: https://ceur-ws.org/Vol-3740/paper-181.pdf.

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

[11] 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.