Science for Fun: The CLEF 2023 JOKER Track on Automatic Wordplay Analysis

Liana Ermakova[1][0000−0002−7598−7474], Tristan Miller[2][0000−0002−0749−1100], Anne-Gwenn Bosser[3], Victor Manuel Palma Preciado[4], Grigori Sidorov[4][0000−0003−3901−3522], and Adam Jatowt[5][0000−0001−7235−0665]

[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
[5] University of Innsbruck, Austria

Abstract Understanding and translating humorous wordplay often requires recognition of implicit cultural references, knowledge of word formation processes, and discernment of double meanings – issues which pose challenges for humans and computers alike. This paper introduces the CLEF 2023 JOKER track, which takes an interdisciplinary approach to the creation of reusable test collections, evaluation metrics, and methods for the automatic processing of wordplay. We describe the track's interconnected shared tasks for the detection, location, interpretation, and translation of puns. We also describe associated data sets and evaluation methodologies, and invite contributions making further use of our data.

Keywords: Wordplay · Puns · Humour · Wordplay interpretation · Wordplay detection · Wordplay generation · Machine translation.

1 Introduction

Humour remains one of the most thorny aspects of intercultural communication. Understanding humour often requires recognition of implicit cultural references or, especially in the case of wordplay, knowledge of word formation processes and discernment of double meanings. These issues raise the question not only of how to translate humour across cultures and languages, but also how to even recognise it in the first place. Such tasks are challenging for humans and computers alike.

The goal of the JOKER track series at the Conference and Labs of the Evaluation Forum (CLEF) is to bring together linguists, translators, and computer scientists in order to create reusable test collections for benchmarking and to explore new methods and evaluation metrics for the automatic processing of wordplay. In the 2022 edition of JOKER (see Section 2), we introduced pilot shared tasks for the classification, interpretation, and translation of wordplay in English and French, and made our data available for an unshared task [9].[6]

[6] In a shared task, the organisers define the evaluation criteria for an open problem in AI and produce a human-annotated data set for training and testing purposes;

For JOKER-2023, we intend to expand the set of languages in our tasks to include Spanish. We also somewhat simplify and streamline the slate of shared tasks, more closely patterning them after the high-level process used by human translators and focusing them on one type of wordplay – puns.

We choose to focus on puns because, despite recent improvements in the quality of machine translation based on machine learning, puns are often held to be untranslatable by statistical or neural approaches [33,1,26]. Punning jokes are a common source of data in computational humour research, in part because of their widespread availability and in part because the underlying linguistic mechanisms are well understood. However, past pun detection data sets [29,39] are problematic because they draw their positive and negative examples from texts in different domains. In JOKER-2023 we attempt to avoid this problem by generating our negative examples by using naïve literal translations, or by slightly editing our positive examples, a technique pioneered by *Unfun.me* [38].

The three shared tasks of JOKER-2023 can be summarised as follows:

1. **Detection and location of puns** in English, French, and Spanish;
2. **Interpretation of puns** in English, French, and Spanish; and
3. **Translation of puns** from English to French and Spanish.

The unshared task of JOKER-2022 saw its data used for a pun generation task potentially aimed at improving interlocutor engagement in dialogue systems. JOKER-2023 will likewise have an unshared task that aims at attracting runs with other, possibly novel, use cases, such as pun generation or humorousness evaluation.

While JOKER-2022 proved to be challenging (with only 13% of evaluated translations being judged successful), this round's larger data set and more constrained, interconnected tasks may present opportunities for better performance.

2 JOKER-2022: Results and Lessons Learnt

Forty-nine teams registered for JOKER-2022, 42 downloaded the data and seven submitted official runs for its shared tasks: nine for Task 1 on classification and interpretation of wordplay [10], four for Task 2 on wordplay translation in named entities [8], and six for Task 3 on pun translation [11]. One additional run was submitted for Task 1 after the deadline. Two runs were submitted for the unshared task, and new classifications were proposed by participants.

Participants' scores on the wordplay classification part of Task 1 were uniformly high, which we attribute to the insufficient expressiveness of our typology and the class imbalance of our data. Due to the expense involved in revising the typology and applying it to new data, we have decided to drop wordplay

task participants then use the publically released training data to develop systems for solving the problem, which the organisers evaluate on the unpublished test data. In an unshared task, the organisers provide annotated data without a particular problem in mind, and participants are invited to use this data to propose and solve novel problems.

classification from JOKER-2023. However, the interpretation part of the task – which required participants to determine both the location and (double) meaning of the wordplay instances – proved to be more challenging, and provoked great interest from the participants. Besides this, we note that providing the location and interpretation of a play on words may be more relevant for downstream processing tasks such as translation [24, p. 86]. For this reason, this part of the task will be repeated in JOKER-2023, albeit with new data.

JOKER-2022's Task 2, on named entity translation, did not see much variety in the participants' approaches, and their low success rates may be due to a lack of context in the data that would be too expensive for us to source. For these reasons, we have opted to discontinue this task for JOKER-2023.

Like Task 1, Task 3 of JOKER-2022 proved to be both popular and challenging, and so we are rerunning it in JOKER-2023 with new data. Task 3 moreover had the side-effect of producing a French-language corpus with positive and negative examples of wordplay, which some participants endeavoured to use for wordplay generation in French (following methods developed for English). The corpus was also reused by the French Association for Artificial Intelligence to organise a jam on wordplay generation in French during a week-long conference [3]. Of particular interest is how humans perceive the generated wordplay. Participants in the jam, for example, raised questions about how to evaluate the humorousness of the system output. Furthermore, a curated selection of sentences generated using our corpus with a large language model[7] was used by some of the present authors during an outreach event, where a public audience was asked to guess if a given humorous sentence was created by an AI or a human. In JOKER-2023, we thus encourage unshared task submissions describing the use of our data for user perception studies and wordplay generation.

3 Shared Tasks

3.1 Task 1: Pun Detection and Location

Description A *pun* is a form of wordplay in which a word or phrase evokes the meaning of another word or phrase with a similar or identical pronunciation [19]. *Pun detection* is a binary classification task where the goal is to distinguish between texts containing a pun and texts not containing a pun. *Pun location* is a finer-grained task, where the goal is to identify which words carry the double meaning in a text known *a priori* to contain a pun.

For example, the first of the following sentences contains a pun where the word *propane* evokes the similar-sounding word *profane*, and the second sentence contains a pun exploiting two distinct meanings of the word *interest*:

[7] The corpus provides numerous instances of particular wellerisms, and thus lends itself well to prompt engineering using large language models. (Wellerisms are a type of humour in which a proverb, idiom, or other well-known saying is subverted, for example by resegmenting it or by reinterpreting it literally.)

(1) When the church bought gas for their annual barbecue, proceeds went from the sacred to the propane.

(2) I used to be a banker but I lost interest.

For the pun detection task, the correct answer for these two instances would be "true", and for the pun location task, the correct answers are respectively "propane" and "interest".

Data The positive examples for Task 1, which will be used for both the detection and location subtasks, consist of short jokes (one-liners), each containing a single pun. These positive examples will be drawn from previously constructed corpora as well as collections that may not have been used in previous shared tasks.[8] In contrast to previously published punning data sets, our negative examples will be generated by the data augmentation technique of manually or semi-automatically editing positive examples in such a way that the wordplay is lost but most of the rest of the meaning remains.[9] In this way, we hope to better minimise the differences in length, vocabulary, style, etc. that were seen in previous pun detection data sets and that could be picked up on by today's neural approaches. Negative examples will be used only for the pun detection subtask.

As usual with shared tasks, data for all tasks will be split into training and test sets, with the training set (including gold-standard labels) published as soon as available, and the test data withheld until evaluation phase.

English. Our training data will include positive examples from the corpora of SemEval-2017 Task 7 [29], SemEval-2021 Task 12 [35], and various other collections. Positive examples in the test data will be drawn, to the extent possible, from jokes not present in past data sets. As mentioned above, negative examples in both the training and test data will be produced by slightly perturbing the positive examples via data augmentation.

French. In 2022, we created a corpus for wordplay detection in French [9,11] based on the translation of the corpus of English puns introduced at SemEval-2017 Task 7 [29]. Some of the translations were machine translations, and others were human translations sourced from a contest or from native francophone students translators. The majority of human translations (90%) preserved wordplay in some form, while only 13% of the machine translations did so. The resulting corpus is homogeneous, across positive and negative examples, in terms of vocabulary

[8] Admittedly, it may be impossible for us to source positive examples that are not discoverable online, unless we pay experienced comedians to produce a large collection of completely novel jokes, which is costly. We will have to rely on participants' good faith that their systems will not detect punning jokes by matching them against a database of web-scraped examples.

[9] The data augmentation technique will be fully described in the task overview paper in the CLEF 2023 proceedings; we hold off on presenting the details here to discourage participants from reverse-engineering it in their classifiers.

and text length, and it maintains the class balance of the original. However, there was an imbalance across the training and test sets with respect to machine vs. human translations, with more machine translations in the test set. This corpus will be improved and extended for use with JOKER-2023. In particular, we will correct the machine vs. human translation imbalance by sourcing additional, manually verified machine translations for the training set. We will also source new positive examples for our test set, and will apply the same data augmentation technique used for our English data.

Spanish. Our Spanish data set is collected from various web sources (blogs, joke compilations, humour forums, etc.) to which we apply the same data augmentation techniques as for the English data.

Evaluation We follow (and thereby facilitate comparison with) SemEval-2017 Task 7 [29] by evaluating pun detection using the precision, recall, accuracy, and F-score measures as used in information retrieval (IR) [25, Section 8.3], and pun location using the corresponding variants of precision, recall, and F-score from word sense disambiguation (WSD) [31].[10]

3.2 Task 2: Pun Interpretation

Description In *pun interpretation*, systems must indicate the two meanings of the pun. The pun interpretation task at SemEval-2017 required systems to annotate the pun with senses from WordNet, and JOKER-2022 expected annotations according to a relatively complex, structured notation scheme. In JOKER-2023, semantic annotations will be in the form of a pair of lemmatised word sets. Following the practice used in lexical substitution data sets [27], these word sets will contain the synonyms (or absent any, the hypernyms) of the two words involved in the pun, excepting any synonyms/hypernyms that happen to share a spelling with the pun as written.[11] This annotation scheme removes the need for participating systems to directly rely on a particular sense inventory or notation scheme.

For example, for the punning joke introduced in Example 1 above, the word sets are {*gas, fuel*} and {*profane*}, and for Example 2, the word sets are {*involvement*} and {*fixed charge*, *fixed cost*, *fixed costs*}.

Data The data will be drawn from the positive examples of Task 1, with the pun word annotated with two sets of words, one for each sense of the pun. Each set of words will contain synonyms or hypernyms of the sense or (in the case of heterographic puns) the latent target word.

[10] The difference between IR-style and WSD-style metrics is that the former require the system to make a prediction for every instance in the data set, whereas the latter do not. IR-style accuracy is equivalent to WSD-style recall.

[11] Synonyms and hypernyms will be sourced preferentially from WordNet (or similar resources for data sets in other languages), and via manual annotation for those words not present in WordNet.

Evaluation Task 2 will be evaluated with the precision, recall, and F-score metrics as used in word sense disambiguation [31], except that each instance will be scored as the average score for each of its senses. Systems need guess only one word for each sense of the pun; a guess will be considered correct if it matches any of the words in the gold-standard set. For example, a system guessing {*fuel*}, {*profane*} would receive a score of 1 for Example 1, and a system guessing {*fuel*}, {*prophet*} would receive a score of 1/2.

3.3 Task 3: Pun Translation

Description The goal of this task is 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 [5,6]. For example, Example 2 might be rendered into French as *J'ai été banquier mais j'en ai perdu tout l'intérêt.* This fairly straightforward translation happens to preserve the pun, since *interest* and *intérêt* share the same ambiguity. Needless to say, this is coincidence does not hold for the majority of punning jokes in our data set (or generally, for that matter).

Data We will provide an updated training and test set of English–French translations of punning jokes, and new sets of English–Spanish ones, similar to English–French data sets we produced for JOKER-2022 [11,9].

Evaluation As we have previously argued [11,9], vocabulary overlap metrics such as BLEU are unsuitable for evaluating wordplay translations. We will therefore continue JOKER-2022's practice of having trained experts manually evaluate system translations according to features such as lexical field preservation, sense preservation, wordplay form preservation, style shift, humorousness shift, etc. and the presence of errors in syntax, word choice, etc. The runs will be 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. We will also experiment with other semi-automatic metrics.

4 State of the Art

Humour is part of social coexistence and therefore is part of interpersonal interactions. This places it in a complicated position, since the perception of humour can be somewhat ambiguous and depends on a number of subjective factors. Thus, dealing with humour, even in its written form, becomes a rather complex undertaking, even for those (computational) tasks that at the first sight seem trivial. Various studies have addressed these tasks, including the detection, classification, and translation of humour, and also determining whether the intention or interpretability of the translated humour is maintained. Some of the

present authors have even designed evaluation campaigns for some of these tasks (e.g., [29,35,10,8,11]), aiming not just to support traditional NLP applications, but also to gain a broader knowledge of the structure and nuances of verbal humour.

Nevertheless, relatively few studies have been carried out on the machine translation (MT) of wordplay. One of the earliest of these [12] proposed a pragmatic-based approach to MT, but no working system was implemented. An interactive method for the computer-assisted translation of puns was recently implemented [24], but it cannot be directly applied for MT. Four teams participated in the pun translation task of JOKER-2022 [7,16,14]; their approaches relied variously on applications of transformer-based models or on DeepL.

Automatic humour recognition has become an emerging trend with the rise of conversational agents and the need for social media analysis [30,15,22,23,34,18,13]. While some systems have achieved decent performance on humour detection, location, and classification tasks [29,10], the lack of high-quality training data has been a limiting factor for further progress, and especially in case of languages other than English [9]. As with translation, many of the JOKER-2022 classification task participants [16,2] favoured applications of large language models such as Google T5 and Jurassic-1.

Other popular application areas in computational humour include humour generation and humorousness evaluation. Recent work in the former area includes template-based approaches for pun generation in English and French [36,20,17], as well as injecting humour into existing non-humorous English texts [37]. Though these tasks have been studied in a monolingual setting, it may be possible to adapt them for a translation task. Work in humorousness evaluation covers methods that attempt to quantify the level of humour in a text, or to rank texts according to their level of humour [40,32,21,4,28]. Such methods also have possible applications in humour translation (e.g., by verifying that a translated joke preserves the level of humour of the original).

5 Conclusion

This paper has described the prospective setup of the CLEF 2023 JOKER track, which features shared tasks on pun detection, location, interpretation, and translation. We will also welcome submissions using our data for other tasks, such as pun generation, offensive joke detection, or humour perception. Please visit the JOKER website at `http://joker-project.com` for further details on the track.

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*.

References

1. Ardi, H., Al Hafizh, M., Rezqi, I., Tuzzikriah, R.: Can machine translations translate humorous texts? Humanus **21**(1) (2022). https://doi.org/10.24036/humanus.v21i1.115698
2. Arroubat, H.: CLEF Workshop: Automatic Pun and Humour Translation Task. In: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (2022)
3. Bosser, A.G., Ermakova, L., Dupin de Saint Cyr Bannay, F., de Loor, P., Charpenay, V., Pépin-Hermann, N., Alcaraz, B., Autran, J.V., Devillers, A., Grosset, J., Hénard, A., Marchal, F.: Poetic or humorous text generation: Jam event at PFIA2022. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) 13th Conference and Labs of the Evaluation Forum (CLEF 2022). pp. 1719–1726. No. Working Notes: JokeR: Automatic Wordplay and Humour Translation in CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (Sep 2022), `https://hal.archives-ouvertes.fr/hal-03795272`
4. Castro, S., Chiruzzo, L., Rosá, A.: Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., de Albornoz, J.C. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages. CEUR Workshop Proceedings, vol. 2150, pp. 187–194. Spanish Society for Natural Language Processing (Sep 2018), `http://ceur-ws.org/Vol-2150/overview-HAHA.pdf`
5. Delabastita, D.: There's a Double Tongue: an Investigation into the Translation of Shakespeare's Wordplay, with Special Reference to Hamlet. Rodopi, Amsterdam (1993)
6. Delabastita, D.: Wordplay as a translation problem: a linguistic perspective. In: Ein internationales Handbuch zur Übersetzungsforschung, vol. 1, pp. 600–606. De Gruyter Mouton (7 2008). https://doi.org/10.1515/9783110137088.1.6.600
7. Dhanani, F., Rafi, M., Tahir, M.A.: FAST_MT participation for the JOKER CLEF-2022 automatic pun and human translation tasks. In: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. p. 14. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (2022)
8. Ermakova, L., Miller, T., Boccou, J., Digue, A., Damoy, A., Campen, P.: Overview of the CLEF 2022 JOKER Task 2: Translate wordplay in named entities. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. CEUR Workshop Proceedings, vol. 3180, pp. 1666–1680 (Aug 2022)
9. Ermakova, L., Miller, T., Regattin, F., Bosser, A.G., Mathurin, E., Corre, G.L., Araújo, S., Boccou, J., Digue, A., Damoy, A., Jeanjean, B.: Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation workshop. In: Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022). LNCS, vol. 13390 (2022)
10. Ermakova, L., Regattin, F., Miller, T., Bosser, A.G., Araújo, S., Borg, C., Corre, G.L., Boccou, J., Digue, A., Damoy, A., Campen, P., Puchalski, O.: Overview of the

CLEF 2022 JOKER Task 1: Classify and explain instances of wordplay. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2022)
11. Ermakova, L., Regattin, F., Miller, T., Bosser, A.G., Borg, C., Jeanjean, B., Mathurin, E., Corre, G.L., Hannachi, R., Araújo, S., Boccou, J., Digue, A., Damoy, A.: Overview of the CLEF 2022 JOKER Task 3: Pun Translation from English into French. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2022)
12. Farwell, D., Helmreich, S.: Pragmatics-based MT and the translation of puns. In: Proceedings of the 11th Annual Conference of the European Association for Machine Translation. pp. 187–194 (Jun 2006), http://www.mt-archive.info/EAMT-2006-Farwell.pdf
13. Francesconi, C., Bosco, C., Poletto, F., Sanguinetti, M.: Error Analysis in a Hate Speech Detection Task: the Case of HaSpeeDe-TW at EVALITA 2018. In: Bernardi, R., Navigli, R., Semeraro, G. (eds.) Proceedings of the 6th Italian Conference on Computational Linguistics (Nov 2018), http://ceur-ws.org/Vol-2481/paper32.pdf
14. Galeano, L.J.G.: LJGG @ CLEF JOKER Task 3: An improved solution joining with dataset from task. In: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. p. 7. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (2022)
15. Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., Rosso, P.: IDAT@FIRE2019: Overview of the track on irony detection in Arabic tweets. In: Proceedings of the 11th Forum for Information Retrieval Evaluation. p. 10–13. Association for Computing Machinery (2019). https://doi.org/10.1145/3368567.3368585
16. Glémarec, L.: Use of SimpleT5 for the CLEF workshop JokeR: Automatic Pun and Humor Translation. In: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. p. 11. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (2022)
17. Glémarec, L., Bosser, A.G., Ermakova, L.: Generating humourous puns in French. In: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022. p. 8. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (2022)
18. Guibon, G., Ermakova, L., Seffih, H., Firsov, A., Le Noé-Bienvenu, G.: Multilingual Fake News Detection with Satire. In: CICLing: International Conference on Computational Linguistics and Intelligent Text Processing. La Rochelle, France (Apr 2019), https://halshs.archives-ouvertes.fr/halshs-02391141
19. Hempelmann, C.F., Miller, T.: Puns: Taxonomy and phonology. In: Attardo, S. (ed.) The Routledge Handbook of Language and Humor, pp. 95–108. Routledge Handbooks in Linguistics, Routledge, New York, NY (Feb 2017). https://doi.org/10.4324/9781315731162-8
20. Hong, B.A., Ong, E.: Automatically extracting word relationships as templates for pun generation. In: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. pp. 24–31. Association for Computational Linguistics, Boulder, Colorado (Jun 2009), https://aclanthology.org/W09-2004
21. Hossain, N., Krumm, J., Gamon, M., Kautz, H.: SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. In: Proceedings of the Fourteenth Workshop

on Semantic Evaluation. pp. 746–758. International Committee for Computational Linguistics (Dec 2020). https://doi.org/10.18653/v1/2020.semeval-1.98
22. Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., Aussenac-Gilles, N.: Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In: 15th Conference of the European Chapter of the Association for Computational Linguistics. vol. 1, pp. 262–272. Association for Computational Linguistics (Apr 2017), https://aclanthology.org/E17-1025.pdf
23. Karoui, J., Farah, B., Moriceau, V., Aussenac-Gilles, N., Hadrich-Belguith, L.: Towards a contextual pragmatic model to detect irony in tweets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. vol. 2, pp. 644–650. Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/P15-2106
24. Kolb, W., Miller, T.: Human–computer interaction in pun translation. In: Hadley, J.L., Taivalkoski-Shilov, K., Teixeira, C.S.C., Toral, A. (eds.) Using Technologies for Creative-Text Translation, pp. 66–88. Routledge (2022). https://doi.org/10.4324/9781003094159-4
25. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
26. Miller, T.: The punster's amanuensis: The proper place of humans and machines in the translation of wordplay. In: Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology. pp. 57–64 (Sep 2019). https://doi.org/10.26615/issn.2683-0078.2019_007
27. Miller, T., Benikova, D., Abualhaija, S.: GermEval 2015: LexSub – A shared task for German-language lexical substitution. In: Proceedings of GermEval 2015: LexSub. pp. 1–9 (Sep 2015)
28. Miller, T., Do Dinh, E.L., Simpson, E., Gurevych, I.: Predicting the humorousness of tweets using Gaussian process preference learning. Procesamiento del Lenguaje Natural **64**, 37–44 (Mar 2020). https://doi.org/10.26342/2020-64-4
29. Miller, T., Hempelmann, C.F., Gurevych, I.: SemEval-2017 Task 7: Detection and interpretation of English puns. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 58–68 (Aug 2017). https://doi.org/10.18653/v1/S17-2005
30. Nijholt, A., Niculescu, A., Valitutti, A., Banchs, R.E.: Humor in human–computer interaction: a short survey. In: Joshi, A., Balkrishan, D.K., Dalvi, G., Winckler, M. (eds.) Adjunct Proceedings: INTERACT 2017 Mumbai. pp. 199–220. Industrial Design Centre, Indian Institute of Technology Bombay (2017), https://www.interact2017.org/downloads/INTERACT_2017_Adjunct_v4_final_24jan.pdf
31. Palmer, M., Ng, H.T., Dang, H.T.: Evaluation of WSD systems. In: Agirre, E., Edmonds, P. (eds.) Word Sense Disambiguation: Algorithms and Applications, chap. 4, pp. 75–106. No. 33 in Text, Speech, and Language Technology, Springer (2007)
32. Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 49–57 (Aug 2017). https://doi.org/10.18653/v1/S17-2004
33. Regattin, F.: Traduction automatique et jeux de mots : l'incursion (ludique) d'un inculte (Mar 2021), https://motsmachines.github.io/2021/en/submissions/Mots-Machines-2021_paper_5.pdf
34. Reyes, A., Rosso, P., Buscaldi, D.: From humor recognition to irony detection: the figurative language of social media. Data & Knowledge Engineering **74**, 1–12 (4 2012). https://doi.org/10.1016/j.datak.2012.02.005

35. Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., Poesio, M.: SemEval-2021 Task 12: Learning with disagreements. In: Proceedings of the 15th International Workshop on Semantic Evaluation. pp. 338–347 (Aug 2021). https://doi.org/10.18653/v1/2021.semeval-1.41
36. Valitutti, A., Toivonen, H., Doucet, A., Toivanen, J.M.: "Let everything turn well in your wife": Generation of adult humor using lexical constraints. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 2, p. 243–248. Association for Computational Linguistics (8 2013), https://aclanthology.org/P13-2044
37. Weller, O., Fulda, N., Seppi, K.: Can humor prediction datasets be used for humor generation? Humorous headline generation via style transfer. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 186–191. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.figlang-1.25
38. West, R., Horvitz, E.: Reverse-engineering satire, or "Paper on computational humor accepted despite making serious advances". In: AAAI'19/IAAI'19/EAAI'19: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. pp. 7265–7272 (Jan 2019). https://doi.org/10.1609/aaai.v33i01.33017265
39. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. p. 2367–2376. Association for Computational Linguistics (9 2015). https://doi.org/10.18653/v1/D15-1284
40. Zhao, Z., Cattle, A., Papalexakis, E., Ma, X.: Embedding lexical features via tensor decomposition for small sample humor recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP). pp. 6376–6381 (Nov 2019). https://doi.org/10.18653/v1/D19-1669