The JOKER Corpus: English–French Parallel Data for Multilingual Wordplay Recognition

Liana Ermakova
HCTI, Université de Bretagne Occidentale
Brest, France
liana.ermakova@univ-brest.fr

Anne-Gwenn Bosser
École Nationale d'Ingénieurs de Brest, Lab-STICC CNRS
UMR 6285
Brest, France
anne-gwenn.bosser@enib.fr

Adam Jatowt
University of Innsbruck
Innsbruck, Austria
jatowt@acm.org

Tristan Miller
Austrian Research Institute for Artificial Intelligence
Vienna, Austria
tristan.miller@ofai.at

ABSTRACT

Despite recent advances in information retrieval and natural language processing, rhetorical devices that exploit ambiguity or subvert linguistic rules remain a challenge for such systems. However, corpus-based analysis of wordplay has been a perennial topic of scholarship in the humanities, including literary criticism, language education, and translation studies. The immense data-gathering effort required for these studies points to the need for specialized text retrieval and classification technology, and consequently for appropriate test collections. In this paper, we introduce and analyze a new dataset for research and applications in the retrieval and processing of wordplay. Developed for the JOKER track at CLEF 2023, our annotated corpus extends and improves upon past English wordplay detection datasets in several ways. First, we introduce hundreds of additional positive examples of wordplay; second, we provide French translations for the examples; and third, we provide negative examples of non-wordplay with characteristics closely matching those of the positive examples. This last feature helps ensure that AI models learn to effectively distinguish wordplay from non-wordplay, and not simply texts differing in length, style, or vocabulary. Our test collection represents then a step towards wordplay-aware multilingual information retrieval.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; **Multilingual and cross-lingual retrieval**; **Test collections**; • **Computing methodologies** → **Language resources**; **Lexical semantics**; *Discourse, dialogue and pragmatics.*

KEYWORDS

annotated datasets for machine learning, wordplay retrieval & analysis, parallel corpora, wordplay detection, wordplay location, text classification

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591885

ACM Reference Format:
Liana Ermakova, Anne-Gwenn Bosser, Adam Jatowt, and Tristan Miller. 2023. The JOKER Corpus: English–French Parallel Data for Multilingual Wordplay Recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591885

1 INTRODUCTION

Wordplay is a common form of humor that can crop up in almost any type of discourse, and for many of these (including domains like literature, advertising, and social conversations) it is actually a recurrent and expected feature. It is therefore important that language technology operating on such discourse types be capable of recognizing and appropriately dealing with instances of wordplay. When it comes to search scenarios, many users prefer reading well-written, astute texts, where subtle language play contributes to forming convincing and persuasive rhetoric. The humorous quality of texts could be treated as one of the cognitive or stylistic characteristics of documents which, alongside comprehensibility, subjectivity, and concreteness, users might expect an information retrieval (IR) system to match [28]. An example of a more specialized IR application is joke retrieval, in which a search engine finds jokes similar in style or content to one provided by the user (typically envisaged as a writer or aficionado of humor) [21, 23]. Despite this, search engines generally do not deal effectively with language that exploits ambiguity or subverts linguistic rules.

While humor and wordplay are widely studied in the humanities and social sciences, they have been largely ignored in information retrieval, including dedicated neural net-based retrieval methods and large language models. This is partly because modern AI tools tend to require quality and quantity of training data that has historically been lacking for humor and wordplay.

In this paper, we introduce the JOKER Corpus, a partly parallel English- and French-language dataset for wordplay detection and location. The JOKER Corpus extends the pun corpora used for shared tasks at SemEval-2017 and SemEval-2021, providing new wordplay translations in French as well as additional, better-quality negative examples of wordplay in both languages. To the best of our knowledge, this makes the JOKER Corpus the first parallel corpus for wordplay in English and French. The corpus has potential uses

in IR research requiring multilingual parallel data, such as cross-language search, but also in applications requiring monolingual data in either language. We also expect the corpus to be useful for research into computational humor and machine translation. The corpus will be published in full under a licence permitting use for research purposes, following an embargo period ending after its use in a CLEF 2023 shared task.

The rest of the paper is organized as follows: Section 2 reviews the state of the art in computational humor and presents an overview of previous humor and wordplay corpora. We introduce the JOKER Corpus, describing the provenance of the original data and our annotation process, in Section 3. An extensive analysis of the corpus and examples of its use for wordplay detection, location, and translation are given in Section 4. The impact of the resource for information retrieval and computational humor research, as well as its potential industrial and societal benefits, are described in Section 5. Section 6 draws conclusions and discusses perspective for future work.

2 STATE OF THE ART

2.1 Computational Approaches to Humor

Recent years have seen an increasing focus on research problems of automatically detecting, generating, or processing verbal humor. Early humor generation techniques tended to involve templates—for example, one system [52] used lexical constraints to produce adult-oriented jokes by changing a single word in a pre-existing text. In another study [24], the authors trained a model to automatically extract humorous templates for use in generating puns. Notwithstanding some creative workarounds [60], the absence of a large pun corpus for training data has hampered efforts to approach pun generation with neural network–based solutions.

The recent popularity of conversational agents and the requirement to handle vast quantities of social media information justify the need for automatic humor detection techniques [37]. For example, humor or irony analysis is important for the development of human-like chatbots, recommender systems, reputation monitoring in social media, and fake news and hate speech detection [19, 22]. Much of the humor detection and interpretation work applicable to these applications has been carried out on humorous wordplay specifically. The earliest evaluation campaign [36] investigated the tasks of **pun detection** (i.e., determining whether or not a text contains a pun), **pun location** (i.e., locating the particular pun word within a text), and **pun interpretation** task (i.e., explaining the double meaning of the pun). The absence of appropriate training data seems to be a significant barrier to further advancements in performance, at least for supervised systems, and in particular for languages other than English.

Recent experiments on the Puns and ShortJokes datasets (introduced in the next section) have demonstrated that using contextualized embeddings is helpful for recognition of humor in general (though not wordplay in particular) [55]. Similarly, satisfactory results have been achieved using a multilingual model based on a pre-trained BERT for Chinese, Russian, and Spanish [54]. Research in automatic humor detection has also targeted tasks in specific domains, such as having Q&A systems recognize humorous questions about products [61].

Relatively little research has focused on machine translation of wordplay. One of the earliest studies on this topic [17] proposed a pragmatics-based approach that accounts for the author's locutionary, illocutionary, and perlocutionary intentions (i.e., the "how", "what", and "why" of the text). However, no implementation of the method was provided. A more recent proposal for an interactive, computer-assisted wordplay translation system [34] was implemented in the PunCAT system [30]. PunCAT skirts the need for a training corpus by relying on the human user to identify plausible translation candidates; in this sense it functions more as an "assisted brainstorming" tool than a machine translation system. While the underlying method is language-independent, the prototype implementation works only for translating wordplay from English into German.

2.2 Existing Humor Corpora

Existing monolingual humor corpora include, *inter alia*, several datasets developed for shared tasks at the International Workshop on Semantic Evaluation (SemEval) [26, 32, 36, 41]. Only a few of these [36, 51] focus specifically on humorous wordplay. The dataset by Mihalcea and Strapparava [33] is comprised of 16,000 humorous and 16,000 non-humorous sentences collected from news titles, proverbs, the British National Corpus, and the Open Mind Common Sense. Yang et al. [58] introduce a dataset that contains 2,400 puns and non-puns from news sources, Yahoo! Answers, and proverbs. The authors prepared negative sentences in a way that minimizes domain differences (i.e., by ensuring similar text length and vocabulary across classes). The Reddit dataset [55] provides jokes segmented into setups and punchlines and rated according to humorousness. The Humicroedit [25] dataset contains news headlines where one word is substituted to evoke incongruity. The authors use the original news headlines as negative (non-humorous) instances. ShortJokes[1] is a collection of 231,657 web-scraped jokes, ranging in length from 10 to 200 characters. While most of the above-mentioned datasets contain only English examples, there are some corpora for other languages including Italian [44], Russian [4, 15], and Spanish [7]. However, to the best of our knowledge, there are no French corpora.

When it comes to wordplay corpora, we are not aware of any previous parallel datasets besides the one created for the previous JOKER shared task at CLEF 2022 [12, 13]. That dataset was created by preparing and creating over a thousand instances of translated wordplay from video games, advertising slogans, literature, and other sources in English and French, mostly consisting of puns and portmanteau-based proper nouns and neologisms. Each sentence was subject to manual investigation and classification into several types of wordplay. Also annotated were the text's lexical-semantic and morphosemantic components. The principal drawbacks of this corpus with respect to wordplay detection and interpretation applications are its lack of negative examples and the complexity of its annotation scheme [11].

[1] https://www.kaggle.com/abhinavmoudgil95/short-jokes

3 RESOURCE

3.1 English Subcorpus

Our English corpus extends those used for the SemEval-2017 shared tasks on pun detection and location [36] and the SemEval-2021 shared task on humor preferences [51] with a number of innovations, as discussed below. The SemEval-2017 data consists of 4,027 unique[2] texts collected from various sources: 2,875 jokes containing a single pun, 520 jokes not containing any pun-based wordplay, and 632 non-humorous texts such as proverbs and aphorisms. To avoid complications associated with the processing of multi-word expressions, the puns were selected such that they contain exactly one content word (i.e., a noun, verb, adjective, or adverb). The data is presented in an XML format, with a custom schema that segments the texts by word tokens and marks up those tokens forming the pun (where present). The SemEval-2021 dataset extends that of SemEval-2017 with 1,000 additional punning jokes from the now-defunct PunOfTheDay.com website, albeit ones that do not always conform to the strict requirements of SemEval-2017.

We have further revised and extended these datasets in three ways. First, we filtered the punning texts from SemEval-2021 to those containing a single pun involving a single content word, leaving us with 209 examples. To these, we added an additional 422 examples, also sourced from PunOfTheDay.com. Second, we have simplified the corpus format by converting the structured XML of SemEval-2017 to a flat tab-delimited text file. This change facilitates visual inspection of the dataset, and also somewhat reduces the barrier to processing it computationally, since users need not rely on (potentially unfamiliar) XML libraries. Our third change addresses the criticisms that positive and negative examples in SemEval-2017 are unbalanced, and that they are too easy to distinguish automatically by dint of their difference in length, lexical choice, and other shallow surface features. To mitigate these issues, we have produced an entirely new set of negative examples that are equal in length to the positive ones, and that very closely match their vocabulary. We have done this by taking about half of the punning jokes and then "ruining" them by substituting a single word (which may be the pun or some other word in the sentence) such that the text remains grammatical and meaningful, but the humorous ambiguity is removed.[3] For example, the punning joke "My insurance did not cover acupuncture, so I got stuck with the bill." was ruined by replacing the word "acupuncture": "My insurance did not cover massage, so I got stuck with the bill." Besides, we applied back translation of puns English↔French to destroy puns. We filtered out all back translations matching the original sentences or containing the punning word annotated for the original puns. Then we manually checked the back translations left in order to split them into puns and non-puns.

The final English portion of the pun detection dataset thus contains 3,506 punning jokes (2,875 from SemEval-2017, 209 from SemEval-2021, and 422 new ones), plus "ruined" versions of 1,708 of these jokes. Though not included in the JOKER Corpus analysis in Section 4 below, we also redistribute the original 1,152 non-puns from SemEval-2017, which users of our dataset may include or discard according to their needs.

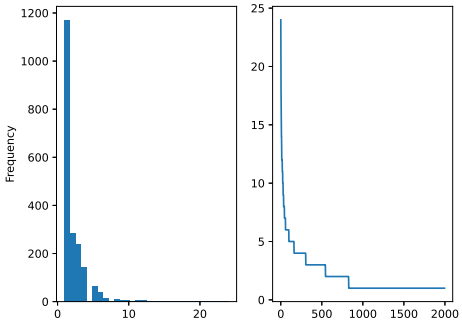

Figure 1: Number of translations per English pun

3.2 French Subcorpus

For the CLEF 2022 JOKER track on wordplay translation, we created a corpus for wordplay detection in French [13, 14] that was based mainly on the English puns from SemEval-2017 [36]. Some of the translations were machine translations, others human translations sourced from a wordplay translation contest or from francophone student translators. All translations were manually annotated with the location of the wordplay, making it, to our knowledge, the first corpus of wordplay locations for French.

However, there is an imbalance across the training and test sets with respect to machine vs. human translations, with more machine translations in the test set [11]. This corpus was improved and extended for the present JOKER Corpus for CLEF 2023. In particular, we corrected the machine vs. human translation imbalance by sourcing additional, manually verified machine translations for the training set and applied the same data augmentation technique used for our English data.

The majority of positive wordplay examples come from human translations, 90% of which preserve the original's wordplay in some form. By contrast, only 13% of the machine translations contain wordplay. Thus, our machine translations serve as our primary source of negative examples. Besides these, we had students translators (all native French speakers) manually produce instances containing wordplay variants, as well as instances of "ruined" wordplay, using essentially the same data augmentation technique as in the English subcorpus. Thus, 10% of non-wordplay was sourced manually, while 33% of wordplay comes from machine translations; this helps ensure that models trained on this data learn to distinguish wordplay from non-wordplay and not machine translations from human ones.

The resulting French subcorpus has 7,306 texts containing wordplay and 9,566 texts without wordplay. As we show in Section 4, the positive and negative examples are homogeneous in terms of vocabulary and text length. As the French data is based on the translation of the English puns, a substantial part of the whole wordplay corpus is parallel, with 6,656 aligned wordplay translations of 2,216 English instances. Figure 1 shows a histogram of the

[2]As originally distributed, the dataset contained 4,030 texts, but three of these were later found to be duplicates [46]. We removed these three duplicates from our corpus.
[3]The general idea behind this technique was pioneered by the Unfun.me and SemEval-2020 Task 7 studies [26, 56].

number of translations per English pun, as well as the count of French translations for each English pun, sorted in descending order. Note that a successful wordplay translation is not necessarily a literal one. For our purposes, we considered a successful translation to be one that (1) preserves the general meaning of the original pun, broadly construed, and (2) contains some sort of wordplay. For example, the punning joke *My name is Wade and I'm in swimming pool maintenance* was successfully translated into French as *Je m'appelle Jacques Ouzy, je m'occupe de l'entretien des piscines.* Here the original pun plays on a proper name (*Wade*) that sounds like the verb for walking through water (*wade*); the French translation similarly uses a plausible proper name (*Jacques Ouzy*) and a water-related homophone (*jacuzzi*). We acknowledge that some contexts might require more strict constraints on the translations, but our translations generally follow the common practices from translation studies.

3.3 Data Format

The resource is provided in two file formats: a tab-delimited text file and a JSON file. For each file format, we provide five files:

- **`joker_detection_en`** Intended for pun detection in English, with the following fields: `id` (a unique identifier), `text` (the text of the instance, which may or may not contain wordplay), and `wordplay` (yes/no)
- **`joker_location_en`** Intended for pun location in English, with the following fields word: `id` (a unique identifier), `text` (the text of the instance, always containing wordplay), and `location` (the portion of the text containing the wordplay)
- **`joker_detection_fr`** Intended for pun detection in French, with the following fields: `id` (a unique identifier), `text` (the text of the instance, which may or may not contain wordplay), and `wordplay` (yes/no)
- **`joker_location_fr`** Intended for pun location in French, with the following fields word: `id` (a unique identifier), `text` (the text of the instance, always containing wordplay), and `location` (the portion of the text containing the wordplay)
- **`joker_parallel`** An index file that maps English instances to their French translations. Fields: `id_en` (unique ID of an English instance), `text_en` (the text of the instance, always containing wordplay in English), `id_fr` (unique ID of a French instance), and `text_fr` (the text of the instance, always containing wordplay in French)

The tab-delimited text file format allows users to easily inspect and edit the corpus in a text editor or spreadsheet. Since the data is pre-organized according to the tasks of wordplay detection, location, and translation, it does not require manual filtering or the requisite know-how to do so, which may be important for users outside of computer science, such as those in the (digital) humanities.

3.4 Licensing and Availability

The individual source texts in our dataset are likely too short to meet the threshold of originality for copyright protection, at least under English law [1]. We have furthermore been advised by the French Technology Transfer Office Ouest Valorisation[4] that research-related distribution of these texts (and their translations and annotations), even if they are indeed copyrightable, would fall within the exceptions provided by article 112–3 of French copyright law (*Code de la propriété intellectuelle*). We, therefore, release our data at no cost under a licence permitting public reuse, modification, and redistribution for research purposes, subject to the attribution requirements recommended by the Office.

The resource will be used for the JOKER shared task[5] at CLEF 2023[6] and will be published in full at the end of the CLEF 2023 evaluation cycle.

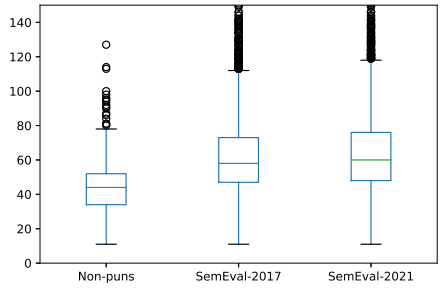

Figure 2: Boxplots of pun and non-pun length in chars from SemEval-2017 and SemEval-2021 data

4 CORPUS ANALYSIS AND USAGE SHOWCASES

4.1 Wordplay Detection

Wordplay detection is a binary text classification task useful for information retrieval, digital humanities, conversational agents, and other humor-aware text processing applications. The classification goal is to determine whether or not a given text contains wordplay. As stated above, previously constructed wordplay datasets suffer from superficial and stylistic differences (in vocabulary, length, etc.) between the positive and negative classes. This is illustrated by the statistics on pun and non-pun text length for the SemEval-2017 and -2021 datasets reported in Table 1, and the corresponding box-and-whisker plot of Figure 2.[7] (The boxes in the figure extend from the first to third quartiles, and are divided at the median. The whiskers show the range of the data no more than 1.5 times the difference in those quartiles, with outliers plotted as separate dots. The y-axis has been capped at 150 characters.) Observe that for both SemEval corpora, the average length of texts containing puns (63 and 65 characters, respectively) is roughly 50% greater than that for non-puns (43 characters). In contrast, in the JOKER Corpus (also Table 1), there is only a 3 or 4 characters' difference in average pun and non-pun length. Our corpus is therefore much more homogeneous in terms of the length of wordplay and non-wordplay instances.

[4] https://www.ouest-valorisation.fr/

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

[6] https://clef2023.clef-initiative.eu/index.php

[7] Note that the SemEval-2021 puns include all the SemEval-2017 puns, while the non-pun data is common to both datasets. We omit from the SemEval-2021 dataset three puns the organizers substituted for the duplicates in SemEval-2017.

Table 1: Statistics on text length, in characters, for the SemEval-2017 and SemEval-2021 corpora

	SemEval-2017	SemEval-2021	SemEval	JOKER			
	puns	puns	non-puns	puns (EN)	non-puns (EN)	puns (FR)	non-puns (FR)
mean	63	65	43	63	59	75	73
std	27	26	14	26	24	31	30
min	11	11	11	11	11	14	16
25%	47	48	34	47	44	57	53
50%	58	60	44	58	54	72	68
75%	73	76	52	74	69	88	86
max	400	400	127	400	400	717	489

Table 2: Lexical overlap between puns and non-puns in datasets

Language	Dataset	COS distance	BLEU
English	SemEval-2017	0.17	5.24
English	SemEval-2021	0.14	14.51
English	JOKER	0.01	64.52
French	JOKER	0.00	59.54

To measure the lexical overlap between wordplay and non-wordplay in each corpus, we calculated the cosine distance (as implemented in the scipy package[8]) between count vector representations[9] as well as the sentence-level BLEU score (as implemented in NLTK[10]). As shown in Table 2, the cosine distance between puns and non-puns is 0.17 and 0.14 for the SemEval-2017 and -2021 datasets, respectively, but is negligible in the JOKER Corpus. JOKER's BLEU score is many times higher than those of the SemEval corpora. Both metrics thus confirm that the JOKER corpus is homogeneous in terms of vocabulary across wordplay and non-wordplay instances.

In order to analyze our resource in terms of its quality for the pun detection task, we compared it with the two previous datasets using the following baselines:

random A naïve baseline that selects a random class for each text.

Ridge [40] The Ridge regression implementation from the sklearn library[11] with TF–IDF vectorization with the parameters `tol = 1e-2, solver = "sparse_cg"`.

NB The Mutlinomial Naïve Bayes implementation from the sklearn library[12] with TF–IDF vectorization.

FastText The FastText implementation[13] from [27].

MLP The MultiLayer Perceptron implementation from the sklearn library[14] with TF–IDF vectorization trained with the parameter `max_iter = 100`.

T5 The SimpleT5[15] implementation of the base T5 model [43] trained with the parameters `source_max_token_len = 100, target_max_token_len = 10, batch_size = 16, max_epochs = 5`. We chose the best model according to the validation loss—i.e., `epoch = 2`.

For all classifiers, we used 80% of shuffled data for training and 20% for testing. We used the default parameters for other settings.

Table 3 reports the classification results (precision, recall, F-score, accuracy, true and false positives, true and false negatives) for the SemEval data. The results suggest that the classifiers, possibly excepting FastText, assign the wordplay label on the basis of text length; the numbers are consistent with the distribution of pun and non-pun text length from Figure 2. The SemEval corpora thus may not be appropriate for train machine learning models for pun detection. In contrast, the classification errors for the JOKER Corpus (see Table 4) do not correlate with text length. As our data is homogeneous in terms of text length and vocabulary, the classification task is more difficult on the JOKER Corpus as it forces models to learn the distinction between wordplay and non-wordplay without relying on irrelevant shallow features. Besides this, for the French corpus, we can observe that the models struggle to distinguish wordplay from non-wordplay and not machine translations from human ones as the errors are different from the distribution of the machine and human translations in the corpus (90% of non-wordplay and 33% of wordplay comes from machine translations).

4.2 Wordplay Location

Wordplay location is a prerequisite for the retrieval of jokes containing a specified punning word. The histogram in Figure 3 shows the percentage of the locations (i.e., positions) of the punning words within the texts of the JOKER Corpus. (In the figure's x-axis, 0 denotes the beginning of the text and 1 its last word.) As it is evident from the plot, wordplay tends to occur towards the end of the text. This placement is common to humor in general.

Table 5 shows the top most frequent punning words in our corpus, along with their corresponding frequencies in the Brown Corpus of Standard American English [20]. The eleven instances with

[8]https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html

[9]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[10]https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[11]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html

[12]https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[13]https://fasttext.cc/docs/en/python-module.html

[14]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[15]https://github.com/Shivanandroy/simpleT5

Table 3: Wordplay detection results for the SemEval corpora

	SemEval-2017								SemEval-2021							
classifier	P	R	F_1	Acc	TP	FP	TN	FN	P	R	F_1	Acc	TP	FP	TN	FN
random	73.66	52.25	61.13	52.36	62.87	66.05	44.46	42.31	74.70	48.37	58.72	48.01	62.89	66.95	41.05	43.52
Ridge	85.94	95.16	90.31	85.36	65.16	49.21	41.89	45.82	85.13	97.53	90.91	85.09	65.30	52.58	39.98	44.29
NB	75.26	99.48	85.69	76.18	64.53	38.33	41.95	43.75	76.75	100.00	86.84	76.84	64.99	—	43.00	42.35
FastText	88.08	88.24	88.16	83.00	66.73	46.81	39.88	51.65	87.99	94.28	91.02	85.79	66.15	45.82	37.07	49.75
MLP	88.51	90.66	89.57	84.86	65.69	51.81	42.26	46.22	87.95	91.16	89.53	83.70	65.72	57.41	41.43	43.74
T5	91.82	95.16	93.46	90.45	65.50	42.57	41.74	49.65	97.49	90.77	94.01	91.15	66.67	48.45	42.13	45.17

Table 4: Wordplay detection results for the JOKER Corpus

	JOKER (EN)								JOKER (FR)							
classifier	P	R	F_1	Acc	TP	FP	TN	FN	P	R	F_1	Acc	TP	FP	TN	FN
random	39.61	48.51	43.61	50.29	63.11	63.53	59.45	61.42	43.92	49.14	46.38	51.35	75.47	75.36	71.82	73.52
Ridge	33.66	25.45	28.98	50.59	59.94	64.48	57.09	67.16	73.76	63.72	68.37	74.76	74.75	76.58	73.03	70.57
NB	31.61	9.08	14.10	56.19	52.59	64.40	59.06	69.49	80.24	49.97	61.58	73.31	73.36	77.47	73.02	68.58
FastText	50.24	30.95	38.31	60.50	64.88	62.63	57.69	71.19	72.49	59.92	65.61	73.10	75.35	75.51	71.66	77.27
MLP	35.56	32.44	33.93	49.94	60.90	64.49	57.34	65.29	68.50	65.38	66.90	72.30	75.70	74.87	72.17	74.16
T5	61.30	75.89	67.82	71.46	64.82	58.63	55.57	70.95	74.78	65.38	69.76	75.73	74.95	76.29	72.41	73.64

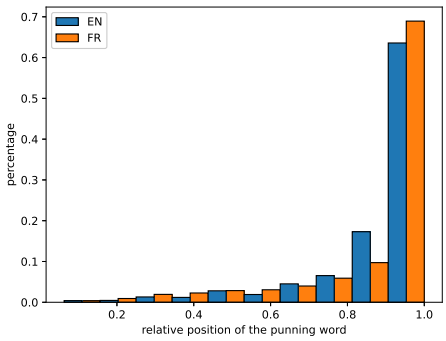

Figure 3: Histogram of the positions of the punning word (location) in the puns in the JOKER corpus

Table 5: The most frequent punning words in English puns with their frequencies in the Brown corpus

punning word	Brown freq.	# of puns in JOKER
point	395	11
shot	27	8
draw	56	8
cut	192	8
still	782	7
downhill	6	6
weight	91	6
son	165	6
steal	5	6
waist	11	6

the most frequent punning word from our corpus, *point*, are as follows:

- I took up teaching fencing as I wanted my students to get the *point*.
- Pencils could be made with erasers at both ends, but what would be the *point*?
- OLD SEAMSTRESSES never die, they just come to the *point*.
- Even the best bird dog is only good to a *point*.
- For a fish, the end of a barbed hook is the "*point*" of no return.
- I used to hate maths but then I realised decimals have a *point*.
- Sometimes a pencil sharpener is needed in order to make a good *point*.
- Why do archers shoot arrows? Could it be they are trying to get a *point* across?
- Decimals have a *point*.
- My friend quit working at the pin factory. He felt there was no *point* to the job.
- I don't understand what the *point* of acupuncture is.

These examples illustrate the potential for our corpus to be used in training IR systems to search for examples of plays on particular words. This functionality can be useful for translators and language educators, as we further discuss in Section 5.

We evaluated the performance of the following baselines for wordplay location:

random The system chooses a word in the text at random.

last The system always predicts the last word in the text. This is a strong baseline, as wordplay is usually placed at the end of the text.

T5 The SimpleT5 implementation of the base T5 model trained with the same parameters as in the wordplay detection task. We chose the best model according to the validation loss—i.e., `epoch = 3`.

Table 6: Wordplay location accuracy

classifier	SemEval-2017	JOKER-EN	JOKER-FR
random	9.90	9.52	6.5
last	50.87	49.15	61.15
T5	96.53	94.46	58.96

Table 7: BLEU scores

model	EN → FR		FR → EN	
	puns	non-puns	puns	non-puns
DeepL	68.90	64.18	26.36	25.82
Opus-MT	63.02	65.96	25.56	25.07
mbart50_m2m	60.57	61.67	25.26	25.05
m2m_100_418M	60.22	61.78	23.94	24.13

Table 6 reports the results for these baselines on the SemEval-2017 and JOKER data. (The SemEval-2021 corpus, as originally distributed, does not contain pun location annotations.) The results suggest that performance is comparable for SemEval-2017 and the English subcorpus of JOKER, but wordplay location in the French subcorpus is more challenging for the large pre-trained AI models. For French data, the T5 model predicted the last word as wordplay location in 44% of cases.

4.3 Wordplay Translation

In order to analyze the JOKER corpus for wordplay translation, we compared the following pre-trained neural machine translation baselines:

DeepL The DeepL machine translation engine[16] integrated into the Phrase (formerly Memsource) cloud-based computer-assisted translation tool[17].

Opus-MT The pre-trained models from Opus-MT [50] implemented in the EasyNMT python package[18].

mbart50_m2m The mBART50 model [49] from Facebook implemented in the EasyNMT package.

m2m_100_418M The M2M 100 model [16] from Facebook with 418 million parameters (1.8 GB) implemented in the EasyNMT package.

Direct application of traditional metrics from machine translation such as BLEU (BiLingual Evaluation Understudy), which measure the vocabulary overlap between the candidate translation and a reference translation [39], is impossible for evaluating wordplay translation quality as they fail to account for meta- or sublexical features such as semantic ambiguity and phonetic similarity. To demonstrate this, we compared BLEU scores for the JOKER subcorpora of wordplay and non-wordplay with respect to translations generated by the state-of-the-art transformer architecture baselines listed above. In Table 7, we report the BLEU scores of the translations from English into French and from French into English. For these, we used the sentence-level BLEU score implementation from the NLTK library[19]. As the number of non-wordplay instances in the English and French subcorpora is larger than the number of wordplay instances, in order to have a fair comparison, we sampled non-wordplay instances equal to the number of references for each language respectively—i.e., 6,656 texts in French and 2,216 texts in English. (Recall that many English puns in our dataset have multiple French translations—see Figure 1.) The BLEU scores for all baselines are almost indistinguishable for translations from French into English (see Table 7), lending credence to our claim that BLEU is inappropriate for evaluating wordplay translation. For the translations from English into French, the situation is even worse, with every baseline, except DeepL, generating translations that are closer to non-wordplay than to wordplay. This underscores the need to develop new metrics applicable to wordplay translation evaluation.

As we showed previously [13], 13% of machine translations are successful seemingly accidentally, owing to the existence of the same word ambiguity in both languages. The existence of these punning words is evident from the analysis of our corpus. Table 8 provides the top 10 pairs of English–French correspondences in punning words. In all cases, the punning words in French are literal translations of the punning words in English, although the cases involving the verbs *take* and *got* seem to be more difficult due to their high ambiguity. Table 9 presents the list of the English punning words producing the highest number of different successful translations in French. Taken together, these results point to the difficulty of cross-lingual retrieval of wordplay; simply translating the source-language punning word into the target language may work in some cases, but in others it will fail to find a large number of relevant examples. The wide variety of translations in our parallel corpus could therefore be used to train such cross-lingual IR models, as well as machine translation systems, to better deal with the multiplicity of wordplay translation strategies.

5 IMPACT

5.1 Wordplay-aware Information Retrieval

Bell [3] reviews the difficulties and potential of using humor and language play in the second-language learning classroom. She notes that being able to understand humor in a foreign language is both a challenge and a common demand from learners, and that the spontaneous use of wordplay in interactions happens only with the most proficient students. On the other hand, some second-language educators have found that puns and plays on words can make texts much more interesting and accessible to students [10]. Several researchers have highlighted the affective benefits of using humorous language play in the classroom, and have even argued for their pedagogical value in emphasizing form [2, 6, 18]. We contend that an information retrieval system capable of recognizing, and perhaps also interpreting, wordplay may be useful for students as a reading aid. An IR system that can both detect wordplay and estimate its difficulty could help educators identify stimulating or problematic instances of wordplay in texts used as language learning materials. The teacher, or a suitably intelligent IR-based system, could then assign or adapt these texts to a given reading

[16] https://www.deepl.com/translator
[17] https://phrase.com/
[18] https://pypi.org/project/EasyNMT/
[19] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

Table 8: Top 10 "easiest" punning words

pun word (EN)	pun word (FR)	# translations	example text (EN)	example translation
faculties	faculté	37	Old school principals never die, they just lose their faculties.	Les vieux proviseurs ne meurent pas, ils perdent toutes leurs facultés.
space	espace	35	Martians welcome. We have space for everyone.	Bienvenue les extraterrestres ! Installez vous, on a créé ces espaces détente pour vous.
root	racine	27	A lot of trees were dying, but they needed to figure out the root of the problem.	De nombreux arbres mouraient mais personne ne trouvait la racine du mal qui les rongeait.
pin	épingler	18	She was suspected of stealing a brooch but they couldn't pin it on her.	Elle s'est fait épingler pour une histoire de broche volée.
deep	profond	18	Well drilling is a deep subject.	Le forage de puits est un sujet profond.
count	compter	17	The inept mathematician couldn't count on his friends.	Un mathématicien qui ne peut compter sur ses amis n'est pas un mathématicien...
take	prendre	17	Doctor, Doctor, what would you take for this cold? - Make me an offer. Next.	Docteur, que prendriez-vous pour un virus ? - Je ne sais pas, faites moi une offre. Suivant.
irrational	irrationnel	16	OLD MATH TEACHERS never die, they just become irrational.	Les vieux profs de maths ne meurent pas, mais ils sont nombreux à devenir irrationnels.
got	prendre	15	A thief who stole a calendar got twelve months.	Un voleur de calendrier vient de prendre douze mois.
burning	brûlant	14	The fire chief was always asked burning questions.	Le chef des pompiers avait l'habitude de poser des questions brûlantes.

Table 9: Top 10 punning words producing the most distinct punning words in French

pun word (EN)	# translations	Pun word in FR translations
cannily	14	cancaner, azurance, boîter, cancana, cancanner, cane, César, conserver, conserve, Cranna, emboîtant, Mostra, ricaner, ricanner
dyed	13	déteinte, cramoisi, marqu, marquer, bac, barbouiller, couleur, couleurectal, coupe, déteindre, fondre, mauvir, pigment
gushed	12	jaillir, jailliser, bouillona, derrick, effusion, epuiser, exploser, mazout, pérorer, pompeux, puiser, raffiné
Avery	11	Élie, cage, chouette, colombe, colombie, corneille, Denis, loiseau, piaf, pigeon, volière
gnus	12	antenne, augnoure, bluff, canard, chouette, gnou, gnous, gnoux, gnue, méduser, oryx, zébu
stairing	11	bâtir, chantier, escalade, fixé, gare, marche, mur, note, pousse, talon, vanne
punctually	15	boucle, crânement, découdre, nasillarde, pénétrant, percer, perçant, piquant, piquer, pointer, pointilleux, pointilleusement, pointu, régulièrement, ponctuellement
orifice	11	bouleau, bourreau, bucau, cabinet, caninet, cariement, fraise, labeurre, mordu, plomb, plomber
drain	11	canaliser, débouché, drainer, drainant, lessiver, pomper, siphonnait, siphonner, tuyau, vanner, vider
draw	11	chapeau, chevalet, dégainer, dégaine, dessinguait, fresque, gouachette, tirer, trace, traceur, vole

proficiency level, thus facilitating the implementation of teaching and learning strategies relying on semantic and pragmatic language play.

The use of wordplay has long been a serious topic of scholarship in literary criticism and analysis. A number of articles and even a few book-length treatises have meticulously catalogued and analyzed the wordplay of individual literary works or their translations, or in the entire œuvres of particular authors [e.g., 9, 29, 38, 45, 57]. Methods for the computational classification and retrieval of wordplay from large text corpora could therefore support humanities research by automating much of the effort in preparing such treatises.

Plays on words are notorious for being particularly difficult to translate, even by skilled human translators. Wordplay in source texts sometimes goes unnoticed by literary translators (who tend to translate into, rather than from, their native languages), and even when it is recognized, translators often eliminate it from the target text, without even adopting any compensatory measures [42]. An information retrieval system for wordplay, therefore, has much to offer the human translator. For one, she could use such a system to automatically retrieve and classify *all* instances of wordplay from the source text she is translating, thus reducing the likelihood of her overlooking or misinterpreting them. (This was one of the functionalities envisaged for Miller's "Punster's Amanuensis" tool [34] and simulated in its prototype implementation; a user study established that this function of the tool was highly prized by human translators [30].) With a similar system, she might query large-scale literary corpora in the target language for wordplay involving a particular word or topic, thereby gathering a set of examples for study and inspiration.

Besides the above-noted uses, a system capable of retrieving texts or individual jokes based on their degree of humorousness would help search users focus on, or eliminate, results according to the desired level of formality or rhetorical effect. While the JOKER Corpus does not presently contain humorousness annotations, adding

them is something we are considering for a future version of the dataset.

5.2 Humor Studies

In addition to IR research, the corpus we present is a valuable resource for the growing body of fundamental research on humor. Indeed, puns and wordplay are frequently used as humorous devices, and research in humor studies is often stymied by the lack of sizeable corpora, as discussed in Section 2. We expect our corpus to be useful in humor research for questions related to the translation of figurative language or the automatic generation of creative and humorous wordplay. We also expect our corpus to be an important step towards advancing research related to the perception of humor by humans, as it can be easily extended in this direction. The perception of humor is known to differ across cultures, which also contributes to challenges in second-language learning, as we remarked above [3]. Work studying how various factors impact the perception of humor often makes use of surveys based on collections of jokes [5, 31]. This is also the case for research on automatic means to predict the humorousness of text [8, 35, 46]. Our corpus, therefore, on the one hand, allows distinguishing humorous wordplay from non-wordplay. On the other hand, it allows for the comparison and analysis of translated jokes, as the majority of our instances in French are translations of English puns.

5.3 Industrial and Societal Impact

A current important issue faced by modern society is the spread of false information and fake news. These can endanger lives, spread hate, and counteract public health messages. On social networks a good joke or pun can help information spread more widely, becoming a meme [48], for instance. Recognizing this, Yeo and McKasy [59] argue that despite often being part of the problem, humor could also be the cure: the emotional flow of news delivery is an important aspect of communication and it could also be useful in debunking misinformation. Information retrieval, machine translation, and language generation tools capable of recognizing or producing wordplay could be useful for detecting and providing counter-measures to fake news propagation. Conversational agents (such as chatbots or ECAs) have also been identified as potential beneficiaries of computational humor and wordplay [37]. Such systems can indeed be used as persuasive agents for coaching applications as the use of humor may decrease reactance [47] and help with persuasive messages [53].

6 CONCLUSION

In this paper, we introduced the JOKER Corpus, a new test and training collection of wordplay, and discussed its usefulness for IR research as well as its potential applications. To the best of our knowledge, it is the first such collection for wordplay detection and location containing French texts. In developing this resource, we resolved some important issues in the English pun corpora previously used for shared tasks at SemEval. The English portion of our corpus almost completely eliminates the lexical differences between puns and non-puns in the SemEval corpora. Our data augmentation technique also allowed us to homogenize data in terms of text length. Taken together, these improvements help ensure that AI models learn to effectively distinguish wordplay from non-wordplay, and not simply texts differing in length, style, or vocabulary. The JOKER Corpus furthermore provides a unique parallel subcorpus of wordplay in English and French. This data could help bridge the gap in cross-lingual wordplay retrieval via machine translation.

Our analysis demonstrated that wordplay detection is still a challenge for state-of-the-art models with $F_1 < 70\%$ for a binary classification problem. This points to the need for wordplay-aware information retrieval systems. The performance of the English pun location baselines is comparable for the SemEval-2017 and JOKER corpora, but the wordplay location in French is more challenging for large pre-trained AI models. In our analysis of the parallel subcorpus, we demonstrated that BLEU is not appropriate to evaluate wordplay translation. This indicates a need to develop new metrics applicable to wordplay translation evaluation.

Besides the JOKER shared task at CLEF, the JOKER Corpus has so far seen internal use by undergraduate computer science students as well as graduate students in humanities. We have formatted it in a way that should be convenient for users with limited technical expertise, such as researchers in the humanities and social sciences. The data is pre-filtered according to the tasks of wordplay detection, location, and translation and so does not require any further processing for these use-cases. The code we used for computing the baselines reported in this paper will be made available as well and can serve as an entry point for how to use the corpus.

The JOKER Corpus should directly contribute to IR research, especially for improving multilingual and cross-language search by taking into account wordplay, therefore helping to address one important shortcoming of current systems. It has applications in fields such as second-language teaching and learning, literary analysis, machine and machine-assisted translation, and humor research. In particular, we expect our corpus to have a direct impact in computational humor research (translation or otherwise) which is a growing and promising area of multidisciplinary research, involving linguistics, psychology, philosophy, and computer science. In the future, we plan to enrich the corpus with annotations for perception annotation by various audiences, which could feed back into the improvement of IR-based systems.

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*. We would like also to thank all colleagues and students who participated in data construction as well as the participants of the translation contest and the CLEF JOKER track.

REFERENCES

[1] BBC 2010. Who, What, Why: Can a Joke Be Copyrighted? BBC News. https://www.bbc.com/news/magazine-10725773

[2] Nancy D. Bell. 2005. Exploring L2 Language Play as an Aid to SLL: A Case Study of Humour in NS–NNS Interaction. *Applied Linguistics* 26, 2 (June 2005), 192–218. https://doi.org/10.1093/applin/amh043

[3] Nancy D. Bell. 2009. Learning About and Through Humor in the Second Language Classroom. *Language Teaching Research* 13, 3 (July 2009), 241–258. https://doi.org/10.1177/1362168809104697

[4] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large Dataset and Language Model Fun-Tuning for Humor Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019). Association for Computational Linguistics, 4027–4032. https://doi.org/10.18653/v1/P19-1394

[5] Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, and Katya Pertsova. 2018. How to Evaluate Humorous Response Generation, Seriously?. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 225–228. https://doi.org/10.1145/3176349.3176879

[6] Cade Bushnell. 2009. "Lego My Keego!": An Analysis of Language Play in a Beginning Japanese as a Foreign Language Classroom. *Applied Linguistics* 30, 1 (March 2009), 49–69. https://doi.org/10.1093/applin/amn033

[7] Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A Crowd-Annotated Spanish Corpus for Humor Analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 7–11. https://doi.org/10.18653/v1/W18-3502

[8] Andrew Cattle and Xiaojuan Ma. 2016. Effects of Semantic Relatedness Between Setups and Punchlines in Twitter Hashtag Games. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. The COLING 2016 Organizing Committee, 70–79. https://aclanthology.org/W16-4308

[9] Anneke de Vries and Adrian J. C. Verheij. 1997. A Portion of Slippery Stones: Wordplay in Four Twentieth-century Translations of the Hebrew Bible. In *Traductio: Essays on Punning and Translation*, Dirk Delabastita (Ed.). St. Jerome, Manchester, 68–94.

[10] Roland Durette. 1971. Notes and News. *The Modern Language Journal* 55, 6 (1971), 382–388.

[11] Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Science for Fun: The CLEF 2023 JOKER Track on Automatic Wordplay Analysis. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13982)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, Berlin, Heidelberg, 546–556. https://doi.org/10.1007/978-3-031-28241-6_63

[12] Liana Ermakova, Tristan Miller, Orlane Puchalski, Fabio Regattin, Élise Mathurin, Sílvia Araújo, Anne-Gwenn Bosser, Claudine Borg, Monika Bokiniec, Gaelle Le Corre, Benoît Jeanjean, Radia Hannachi, Ġorġ Mallia, Gordan Matas, and Mohamed Saki. 2022. CLEF Workshop JOKER: Automatic Wordplay and Humour Translation. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Lecture Notes in Computer Science, Vol. 13186. Springer International Publishing, Cham, 355–363. https://doi.org/10.1007/978-3-030-99739-7_45

[13] Liana Ermakova, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, Albin Digue, Aurianne Damoy, and Benoît Jeanjean. 2022. Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022) (Lecture Notes in Computer Science, Vol. 13390)*, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, Cham, 447–469. https://doi.org/10.1007/978-3-031-13643-6_27

[14] Liana Ermakova, Fabio Regattin, Tristan Miller, Anne-Gwenn Bosser, Claudine Borg, Benoît Jeanjean, Élise Mathurin, Gaëlle Le Corre, Radia Hannachi, Sílvia Araújo, Julien Boccou, Albin Digue, and Aurianne Damoy. 2022. Overview of the CLEF 2022 JOKER Task 3: Pun Translation from English into French. 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, Vol. 3180)*, Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast (Eds.). 1681–1700.

[15] Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. 2018. Stierlitz Meets SVM: Humor Detection in Russian. In *Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018 (Communications in Computer and Information Science, Vol. 930)*, Dmitry Ustalov, Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka (Eds.). Springer, Cham, Switzerland, 178–184. https://doi.org/10.1007/978-3-030-01204-5_17

[16] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-centric Multilingual Machine Translation. *The Journal of Machine Learning Research* 22, 1 (Jan. 2021), 4839–4886.

[17] David Farwell and Stephen Helmreich. 2006. Pragmatics-based MT and the Translation of Puns. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*. 187–194. http://www.mt-archive.info/EAMT-2006-Farwell.pdf

[18] Ross Forman. 2011. Humorous Language Play in a Thai EFL Classroom. *Applied Linguistics* 32, 5 (Dec. 2011), 541–565. https://doi.org/10.1093/applin/amr022

[19] Chiara Francesconi, Cristina Bosco, Fabio Poletto, and Manuela Sanguinetti. 2018. Error Analysis in a Hate Speech Detection Task: The Case of HaSpeeDe-TW at EVALITA 2018. In *Proceedings of the 6th Italian Conference on Computational Linguistics*, Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro (Eds.). http://ceur-ws.org/Vol-2481/paper32.pdf

[20] W. N. Francis and H. Kucera. 1979. *Brown Corpus Manual*. Technical Report. Department of Linguistics, Brown University, Providence, RI, USA. http://icame.uib.no/brown/bcm.html

[21] Lisa Friedland and James Allan. 2008. Joke Retrieval: Recognizing the Same Joke Told Differently. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 883–892. https://doi.org/10.1145/1458082.1458199

[22] Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2019. Multilingual Fake News Detection with Satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*. La Rochelle, France. https://halshs.archives-ouvertes.fr/halshs-02391141

[23] Dhruv Gupta, Mark Digiovanni, Hiro Narita, and Ken Goldberg. 1999. Jester 2.0 (Demonstration Abstract): Collaborative Filtering to Retrieve Jokes. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, California, USA) *(SIGIR '99)*. Association for Computing Machinery, New York, NY, USA, 333. https://doi.org/10.1145/312624.312770

[24] Bryan Anthony Hong and Ethel Ong. 2009. Automatically Extracting Word Relationships as Templates for Pun Generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*. Association for Computational Linguistics, 24–31.

[25] Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President Vows to Cut Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 133–142. https://doi.org/10.18653/v1/N19-1012

[26] Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. In *Proceedings of the 14th Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, 746–758. https://doi.org/10.18653/v1/2020.semeval-1.98

[27] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 427–431. https://aclanthology.org/E17-2068

[28] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Investigating Users' Query Formulations for Cognitive Search Intents. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 577–586. https://doi.org/10.1145/2600428.2609566

[29] Stefan Daniel Keller. 2009. *The Development of Shakespeare's Rhetoric: A Study of Nine Plays*. Number 136 in Swiss Studies in English. Narr, Tübingen.

[30] Waltraud Kolb and Tristan Miller. 2022. Human–Computer Interaction in Pun Translation. In *Using Technologies for Creative-Text Translation*, James Luke Hadley, Kristiina Taivalkoski-Shilov, Carlos S. C. Teixeira, and Antonio Toral (Eds.). Routledge, 66–88. https://doi.org/10.4324/9781003094159-4

[31] Tiffany J. Lawless, Conor J. O'Dea, Stuart S. Miller, and Donald A. Saucier. 2020. Is It Really Just a Joke? Gender Differences in Perceptions of Sexist Humor. *HUMOR* 33, 2 (May 2020), 291–315. https://doi.org/10.1515/humor-2019-0033

[32] J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (2021-08). Association for Computational Linguistics, 105–119. https://doi.org/10.18653/v1/2021.semeval-1.9

[33] Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 531–538. https://doi.org/10.3115/1220575.1220642

[34] Tristan Miller. 2019. 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*. 57–64. https://doi.org/10.26615/issn.2683-0078.2019_007

[35] Tristan Miller, Erik-Lân Do Dinh, Edwin Simpson, and Iryna Gurevych. 2019. OFAI–UKP at HAHA@IberLEF2019: Predicting the Humorousness of Tweets

Using Gaussian Process Preference Learning. In *Proceedings of the Iberian Languages Evaluation Forum (CEUR Workshop Proceedings, Vol. 2421)*, Miguel Ángel García Cumbreras, Julio Gonzalo, Eugenio Martínez Cámara, Raquel Martínez Unanue, Paolo Rosso, Jorge Carrillo de Albornoz, Soto Montalvo, Luis Chiruzzo, Sandra Collovini, Yoan Guitiérrez, Salud Jiménez Zafra, Martin Krallinger, Manuel Montes y Gómez, Reynier Ortega-Bueno, and Aiala Rosá (Eds.). 180–190.

[36] Tristan Miller, Christian F. Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. 58–68. https://doi.org/10.18653/v1/S17-2005

[37] Anton Nijholt, Andreea Niculescu, Alessandro Valitutti, and Rafael Enrique Banchs. 2017. Humor in Human–Computer Interaction: A Short Survey. In *Proceedings of INTERACT 2017*.

[38] Malcolm Offord. 1997. Mapping Shakespeare's Puns in French Translations. In *Traductio: Essays on Punning and Translation*, Dirk Delabastita (Ed.). St. Jerome, Manchester, 234–260.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318. https://doi.org/10.3115/1073083.1073135

[40] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[41] Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 49–57. https://doi.org/10.18653/v1/S17-2004

[42] Marlies Gabriele Prinzl. 2016. Death to Neologisms: Domestication in the English Retranslations of Thomas Mann's *Der Tod in Venedig*. *International Journal of Literary Linguistics* 5, 3 (2016). https://doi.org/10.15462/ijll.v5i3.73

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[44] Antonio Reyes, Davide Buscaldi, and Paolo Rosso. 2009. An Analysis of the Impact of Ambiguity on Automatic Humour Recognition. In *Text, Speech and Dialogue*, Václav Matoušek and Pavel Mautner (Eds.). Springer, Berlin, Heidelberg, 162–169. https://doi.org/10.1007/978-3-642-04208-9_25

[45] Frankie Rubinstein. 1984. *A Dictionary of Shakespeare's Sexual Puns and Their Significance*. Macmillan, London.

[46] Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5716–5728. https://doi.org/10.18653/v1/P19-1572

[47] Paul Skalski, Ron Tamborini, Ed Glazer, and Sandi Smith. 2009. Effects of Humor on Presence and Recall of Persuasive Messages. *Communication Quarterly* 57, 2 (May 2009), 136–153. https://doi.org/10.1080/01463370902881619

[48] Viriya Taecharungroj and Pitchanut Nueangjamnong. 2015. Humour 2.0: Styles and Types of Humour and Virality of Memes on Facebook. *Journal of Creative Communications* 10, 3 (Nov. 2015), 288–302. https://doi.org/10.1177/0973258615614420 Publisher: SAGE Publications India.

[49] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv preprint arXiv:2008.00401* (2020). https://doi.org/10.48550/arXiv.2008.00401

[50] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building Open Translation Services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 479–480. https://aclanthology.org/2020.eamt-1.61

[51] Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 Task 12: Learning with Disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation*. 338–347. https://doi.org/10.18653/v1/2021.semeval-1.41

[52] Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2013. "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. Association for Computational Linguistics, 243–248. https://aclanthology.org/P13-2044

[53] Nathan Walter, Michael J Cody, Larry Zhiming Xu, and Sheila T Murphy. 2018. A Priest, a Rabbi, and a Minister Walk into a Bar: A Meta-Analysis of Humor Effects on Persuasion. *Human Communication Research* 44, 4 (Oct. 2018), 343–373. https://doi.org/10.1093/hcr/hqy005

[54] Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified Humor Detection Based on Sentence-pair Augmentation and Transfer Learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 53–59. https://aclanthology.org/2020.eamt-1.7

[55] Orion Weller and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3621–3625.

[56] Robert West and Eric Horvitz. 2019. Reverse-Engineering Satire, or "Paper on Computational Humor Accepted Despite Making Serious Advances". In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 7265–7272. https://doi.org/10.1609/aaai.v33i01.33017265

[57] Leopold Wurth. 1895. *Das Wortspiel bei Shakspere*. Wilhelm Braumüller, Vienna.

[58] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2367–2376. https://doi.org/10.18653/v1/D15-1284

[59] Sara K. Yeo and Meaghan McKasy. 2021. Emotion and Humor as Misinformation Antidotes. *Proceedings of the National Academy of Sciences* 118, 15 (April 2021), e2002484118. https://doi.org/10.1073/pnas.2002484118

[60] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics, 1650–1660. https://doi.org/10.18653/v1/P18-1153

[61] Yftah Ziser, Elad Kravi, and David Carmel. 2020. *Humor Detection in Product Question Answering Systems*. Association for Computing Machinery, New York, NY, USA, 519–528. https://doi.org/10.1145/3397271.3401077