KoWit-24: A Richly Annotated Dataset of Wordplay in News Headlines

Alexander Baranov
HSE University
ambaranov@hse.ru

Anna Palatkina
Interdisciplinary Transformation University Austria
anna.palatkina@it-u.at

Yulia Makovka
Independent Researcher
yulya.makovka@yandex.ru

Pavel Braslavski
HSE University
Ural Federal University
pbras@yandex.ru

Abstract

We present KoWit-24, a dataset with fine-grained annotation of wordplay in 2,700 Russian news headlines. KoWit-24 annotations include the presence of wordplay, its type, wordplay anchors, and entities the wordplay refers to. Unlike the majority of existing collections of *canned* jokes, KoWit-24 provides wordplay *contexts* – each headline is accompanied by the news lead and summary. The most common type of wordplay in the dataset is the transformation of collocations, idioms, and named entities – the mechanism that has been underrepresented in previous humor datasets. Our experiments with five LLMs show that there is ample room for improvement in wordplay detection and interpretation tasks. The dataset and evaluation scripts are available at https://github.com/Humor-Research/KoWit-24.

1 Introduction

Wordplay refers to creative language use that often purposely violates the linguistic norms and aims to draw attention, entertain, and amuse the reader. This umbrella term incorporates various techniques, such as punning, spoonerism, oxymoron, portmanteau, and their combinations. Even most modern advanced LLMs fail to recognize humor and struggle with generating original jokes (Ismayilzada et al., 2024). At the same time, sense of humor is a desirable trait of conversational agents (Shin et al., 2023).

A play on words is quite frequent in news (Partington, 2009; Monsefi and Sepora, 2016); see an example in Figure 1. In this paper, we present KommersantWit (KoWit-24), a collection of headlines from the Russian business daily Kommersant that is known for its distinctive ironic style. The total size of the dataset is 2,700 headlines, about half of which are annotated as containing wordplay. Each wordplay-bearing headline is assigned up to two wordplay mechanisms from a set of eight and has an annotated *anchor* (wordplay-triggering word or phrase). In addition, we provide a word, phrase, or entity the wordplay makes reference to along with a Wikipedia/Wiktionary link, if possible. Importantly, wordplay examples in KoWit-24 are contextualized: each headline is accompanied by a short description of the news story (lead) and a summary.

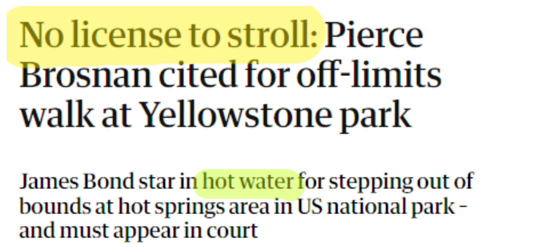

Figure 1: Wordplay example from The Guardian. The part highlighted in yellow refers to *Licence to Kill*, a concept popularized in the James Bond universe and the eponymous film, while the phrase in green allows both idiomatic and literal readings in this context. Source: https://bit.ly/wpbrosnan

KoWit-24 has several features that distinguish it from other humor datasets: 1) associated contexts, 2) a large proportion of transformation-based wordplay examples underrepresented in the previous datasets, 3) non-English content, 4) multi-level annotation, and 5) composition: items with and without wordplay come from the same source.

We conducted wordplay detection and interpretation experiments based on KoWit-24 using a representative set of five LLMs. The results show that there is room for improvements even for GPT-4o, a definitive leader in both tasks.

2 Related Work

In their pioneering paper, Mihalcea and Strapparava (2005) presented a dataset containing 16k one-

Proceedings of Recent Advances in Natural Language Processing, pages 125–132
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-015

liners collected online and an equal number of non-humorous sentences. Since then, several similar datasets have been released, including those that use *reddit* as a source for humorous texts (Yang et al., 2015; Chen and Soo, 2018; Weller and Seppi, 2020; Tang et al., 2023). An alternative approach involves human editing: West and Horvitz (2019) designed an online game in which participants had to edit satirical headlines from *TheOnion* to make them unfunny, while Hossain et al. (2019, 2020b) explored the opposite direction: volunteers and crowd workers had to make news headlines funny with minimal editing. Several SemEval shared tasks have produced new datasets and sparked broader interest in computational humor (Potash et al., 2017; Hossain et al., 2020a; Meaney et al., 2021). Baranov et al. (2023) provide in-depth analysis of existing humor datasets.

While the majority of the datasets contain binary labels or funniness scores, a few provide more detailed annotations. EnglishPuns (Miller et al., 2017) contains annotations of pun type and punning words along with their WordNet senses. Zhang et al. (2019) annotated a collection of Chinese jokes with keywords, character roles, place, humor category, and funniness score. EnglishPuns also became the basis for the ExPUN dataset (Sun et al., 2022), which additionally contains understandability, offensiveness, and funniness scores, as well as keywords important for understanding the joke and natural language explanations.

Most humor-related datasets are in English, but there are also datasets for Italian (Buscaldi and Rosso, 2007), Spanish (Castro et al., 2018), and Portuguese (Inacio et al., 2024). The Russian FUN dataset (Blinov et al., 2019) contains more than 150k funny short texts collected online and the same number of non-humorous forum posts. JOKER (Ermakova et al., 2023) is a rare example of a bilingual collection: it extends EnglishPuns with French translations.

A study by Xu et al. (2024) is close to ours: they evaluate pun detection, explanation, and generation abilities of LLMs using English ExPUN dataset.

3 Methodology

3.1 Data Collection

Kommersant is a Russian news outlet with both print and web editions.[1] Founded in 1990, the newspaper is one of the main Russian business dailies. Since its inception, Kommersant has developed its own distinctive ironic and playful style, which is best reflected in its headlines (Khazanov, 2023; Chernyshova, 2021; Tymbay, 2024).

We collected data from Kommersant via its RSS feed[2] during the period from Jan 2021 to Dec 2023. Each data item corresponds to an article on the website and has the following fields: URL, category (World news, Business, etc.), headline, lead, summary, timestamp, and an optional image link.

3.2 Wordplay Definition

Wordplay is a multifaceted and somewhat ambiguous concept. As the theoretical foundation for our work, we adopted the conceptual framework introduced and applied to the analysis of wordplay in a large collection of British news headlines by Partington (2009). Partington defines **two interpretations** associated with the text as the main characteristics of wordplay. In addition, these two meanings must be somewhat *opposed*, and the wordplay must be deliberately constructed. Partington distinguishes two main wordplay mechanisms: 1) *relexicalization* and 2) *reworking/reconstruction*. The former corresponds to traditional puns, where two meanings arise from either lexical ambiguity (homonymic puns) or phonetic ambiguity (homophonic puns). In the case of reworking/reconstruction, wordplay is based on the modification (*reworking*) of a known phrase; its effect lies in the interplay of the meanings of the present phrase and the original one that the hearer/reader *reconstructs*. Partington points out that wordplay often involves different kinds of preconstructed *phrases*, such as proverbs, quotations, idioms, common collocations, film and book titles, etc. Note that this wordplay definition differs, for example, from those of Monsefi and Sepora (2016) and Brugman et al. (2023). In these studies linguistic devices such as personification, metaphor, metonymy, etc. in news headlines are attributed to wordplay.

3.3 Data Annotation

At the base of KoWIT-24 is the binary annotation of the wordplay presence. For the headlines with

[1] `https://www.kommersant.ru/about` (in Russian)

[2] `https://www.kommersant.ru/RSS/news.xml`; Kommersant grants permission to use its materials, provided that no more than 30% of the original article is used, the text remains unaltered and an attribution is given, see `https://www.kommersant.ru/copyright` (in Russian). The collected data complies with these requirements.

WP type	Original/transliterated headline and translated lead	Literal translation and interpretation
Polysemy	«Волгу» не могут заставить течь быстрее Volgu ne mogut zastavit' tech' bystree The speed limit on the M7 federal highway in the Moscow region remains unchanged	*"Volga" cannot be forced to flow faster.* Volga can both refer to the Volga river and federal highway "Volga".
Homonymy	Туризм подрастерял Шарм Turizm podrasteryal Sharm Operators adjust their Egyptian programs	*Tourism has lost its charm.* Russian *Sharm* can also refer to a shortened form of *Sharm El Sheikh*, a holiday resort in Egypt.
Homophony	Из-под земли до стали Iz-pod zemli do stali The mineral extraction tax for metallurgical companies will be increased starting in 2022	The headline sounds like an idiom *Iz-pod zemli dostali*, literally *Got out from under the ground*, equivalent to English *Left no stone unturned*, but the spelling of the last word (*dostali* → *do stali*) allows for a different reading: *From under the ground to the <u>steel</u>.*
Collocation	Особо бумажные персоны Osobo bumazhnye persony How private investors are reshaping Russia's IPO market	*Very paper persons* refers to *Very important persons*; Russian original/substitute words (*vazhnye/bumazhnye*) rhyme.
Idiom	Код накликал / Kod naklikal Why and for whom open-source software matters in Russia	*Code clicked* sounds in Russian similar to *Cat cried*, an idiom describing a very small amount of something.
Reference	Миссия сократима / Missiya sokratima How Russia could respond to the expulsion of its diplomats from NATO's Brussels mission	*Mission reducible* refers to the *Mission: Impossible* film series.
Nonce word	От запчастного к общему Ot zapchastnogo k obshchemu Car deficit sends Russian sales into reverse	*From sparepartish to a whole*: the headline refers to the induction principle *from a part to a whole* with a neologism adjective derived from the noun *zapchast'* (*spare part*).
Oxymoron	Новый премьер Израиля начал со старого Noviy premier Izrailya nachal so starogo Naftali Bennett heads to Washington	*<u>New</u> Israeli PM started with <u>old</u>* (tricks).

Table 1: Original wordplay examples along with interpretations.

wordplay, we provide further annotations: 1) wordplay type, 2) *anchors*, i.e. words or phrases that trigger the wordplay, 3) *anchor reference*, e.g. a similar-sounding word or original phrase the anchor refers to (note that there is no reference in case of homographic puns that are based on polysemous words), 4) for headlines that are modifications of a collocation, an idiom or refer to a popular entity (such as movie or book titles, catchphrases, etc.), we provide a corresponding link to Wiktionary or Wikipedia, if possible.

The annotation was done using the Label Studio tool[3] by three authors of the paper, two of whom are professional linguists and one is a computer scientist; all three are Russian native speakers and have an extensive experience with NLP-related projects. Translated annotation guidelines can be found in the repository.

Three annotators labeled each element of the data in parallel, making notes on ambiguous cases that were later discussed. We compared the results, discussed discrepancies, and reconciled them in the annotation process. The average of three pairwise Cohen's kappas for the initial wordplay annotations *before discussion* was 0.42, indicating the non-trivial nature of the task (two annotators with linguistic background showed better agreement with $\kappa = 0.58$).[4] The majority of discrepancies were found in the Reference and Collocation categories. There was virtually no disagreement when annotating Oxymoron and Homonymy. However, we hope that subsequent reconciliation of discrepancies ensures a high quality of the resulting annotation. The overlapping annotation was particularly useful: the different cultural preferences and backgrounds of the annotators allowed to get a higher coverage, as not all interpretations are obvious and immediately understandable.

Later, we assigned up to two mechanisms to

[3] https://labelstud.io/

[4] Low agreement is typical for humor-related annotation. E.g., Sun et al. (2022) reported Cohen's kappa values of 0.58 for joke definitions and 0.41 for joke ratings. Similar low agreement has been reported in sarcasm (Oprea and Magdy, 2020) and insult (Mathew et al., 2021) detection studies.

	Wordplay type	*#*	*AAL*	*Links*
Puns	Polysemy	190	1.51	
	Homonymy	26	1.57	
	Phonetic similarity	98	1.80	
Trans.	Collocation	423	2.64	126
	Idiom	177	3.43	118
	Reference	353	3.73	214
	Nonce word	185	1.44	
	Oxymoron	48	2.02	

Table 2: Wordplay types, average anchor length (AAL) in words, and wiki links in KoWit-24. Three mechanisms at the top of the table correspond to traditional *puns*. Three mechanisms in the middle are based on *transformations* of existing phrases. Note that some items are assigned two mechanisms, so the sum of the counts exceeds the number of headlines with wordplay in the dataset (1,340). The last column shows the number of wiki links for transformation-based types.

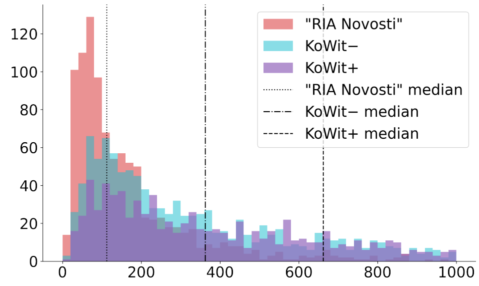

Figure 2: Perplexity distribution of the headlines in two (+/–) KoWit-24 classes and RIA Novosti collection. Vertical lines correspond to the medians of the distributions (note that the histogram is truncated).

the wordplay headlines identified in the first phase. The approach was mainly data-driven: we grouped the headlines based on the similarity of their wordplay mechanisms as we went through the collection, assigned labels, and occasionally re-annotated some items. The final list of the wordplay mechanisms used in the annotation is given in Table 1 along with examples. The two main groups, *Puns* and *Transformations* (see Table 2), correspond to the aforementioned mechanisms of Partington (2009) – *relexicalization* and *reworking/reconstruction*. Within puns, we distinguish homonymic puns based on either on *polysemy* or *homonymy*, as well as homophonic puns based on *phonetic similarity*. We classify transformation-based puns by their source phrases – common *collocations*, *idioms*, or *references* – named entities such as quotations, book or film titles, etc. *Nonce words* (occasionalisms) and *oxymorons* form a separate group.

Finally, we annotated wordplay anchors, provided anchor references, and, if possible, added a link to the corresponding Wiktionary or Wikipedia page.[5] An example of a dataset record is available in our GitHub repository.[6]

3.4 Dataset Statistics and Analysis

In total, we annotated 2,700 headlines, of which 1,340 contained wordplay, so the dataset is almost perfectly balanced. It is interesting to note that the wordplay headlines are on average one word shorter than their counterparts (3.88 vs. 4.81 words). To characterize KoWit-24, we calculated perplexity of the headlines in both classes using ruGPT-3.5 13B model (SaluteDevices team, 2023) and juxtaposed them with the news headlines from the RIA Novosti dataset (Gavrilov et al., 2019). As Figure 2 shows, Kommersant headlines have higher perplexity than the more reserved headlines of the state-owned agency RIA Novosti, and the KoWit-24 headlines with wordplay deem even more ‘unusual’ than their counterparts.

Distribution of headlines by wordplay type can be seen in Table 2. The most frequent wordplay mechanism in our dataset appeared to be the modification of existing well-known phrases – collocations, idiomatic expressions, or named entities. Notably, this type of wordplay is barely presented in previous humor datasets. As anticipated, the average anchor lengths are also higher in the transformation-based classes (average anchor length in the whole collection is 2.65 words). Our observations are in good agreement with the study by Partington (2009), who argues that the wordplay occurs mainly at the *phrasal* level.

About half of all headlines with transformation-based wordplay are provided with Wikipedia (290) and Wiktionary (168) links pointing to descriptions of the source phrases/entity names. The presence of an article in one of the wikis can be seen as an indicator of the popularity of the original phrase/entity, which reduces the risk of subjective and spurious associations.

4 Experiments

For the experiments, we allocated 200 records (100 from each class) for the development set, making sure that all wordplay types were repre-

[5] Wordplay examples in Figure 1 would be annotated with *Reference* and *Polysemy* types, respectively. The highlighted spans would be annotated as wordplay anchors, with *Licence to Kill* as the anchor reference accompanied by the corresponding Wikipedia link.

[6] https://github.com/Humor-Research/KoWit-24

Wordplay type	#	*GigaChat Lite*		*GigaChat Max*		*YaGPT4*		*GPT-4o*	
		simple	extended	simple	extended	simple	extended	simple	extended
Polysemy	168	0.56	0.74	0.57	0.57	0.05	0.23	**0.88**	0.86
Homonymy	22	0.50	0.59	0.50	0.64	0.14	0.23	0.68	**0.82**
Phonetic similarity	88	0.40	0.74	0.44	0.58	0.10	0.15	0.81	**0.90**
Collocation	393	0.47	0.70	0.48	0.58	0.09	0.20	0.78	**0.87**
Idiom	164	0.50	0.74	0.49	0.65	0.12	0.38	0.87	**0.96**
Reference	326	0.49	0.71	0.44	0.58	0.06	0.23	0.76	**0.85**
Nonce word	166	0.52	0.81	0.45	0.60	0.21	0.27	0.87	**0.96**
Oxymoron	34	0.68	0.79	0.68	0.71	0.21	0.41	**0.85**	0.82
Total	1,361	0.50	0.72	0.48	0.59	0.10	0.24	0.81	**0.88**

Table 3: Recall on wordplay detection by type with simple/extended prompt (Mistral's all-zero scores are not shown).

Wordplay type	#	*GigaChat Lite*		*GigaChat Max*		*YaGPT4*		*Mistral*		*GPT-4o*	
		manual	auto	manual	auto	manual	auto	manual	auto	manual	auto
Polysemy	12	0.17	0.33	0.33↑	0.25	0.08	0.00	0.17	0.33	**0.50**	0.50
Homonymy	7	0.14	0.43	0.29	0.29	0.14	0.14	0.00	0.43	**0.43**↑	0.29
Phonetic similarity	85	0.11	0.25	0.28	0.39	0.11	0.21	0.15	0.32	**0.52**↑	0.51
Collocation	393	0.07	0.18	0.27	0.27	0.16	0.20	0.20	0.30	**0.44**↑	0.41
Idiom	164	0.15	0.18	0.37↑	0.30	0.32↑	0.28	0.34↑	0.30	**0.55**↑	0.48
Reference	326	0.10	0.12	0.25↑	0.23	0.20↑	0.16	0.23	0.25	**0.46**↑	0.36
Nonce word	166	0.15	0.46	0.29	0.57	0.28	0.43	0.28	0.57	**0.61**	0.69
Oxymoron	6	0.67	0.67	0.50	0.50	0.33	0.33	**0.83**	0.83	0.67↑	0.50
Total	1,159	0.11	0.19	0.28	0.28	0.20	0.22	0.24	0.30	**0.48**↑	0.43

Table 4: Recall on wordplay interpretation by type; manual and automatic evaluation (↑ marks cases, where manual score exceeds the automatic one).

sented. Thus, the test set contains 2,500 headlines (1,290 with and 1,310 without wordplay). Since the dataset is primarily intended for experiments with modern LLMs in few- or zero-shot mode, we didn't allocate a dedicated training set. With 2k+ test items, KoWIT-24 should ensure low variance in repeated runs.

We experimented with five LLMs and two tasks – 1) wordplay detection and 2) wordplay interpretation. The five LLMs are a representative mix of open/closed, medium-sized/large, and Russian-centric/multilingual models. Details about the models are provided in the extended arXiv version of this paper.[7]

When using LLMs, the temperature was set to 0.1 for the GPT-4o, GigaChat Lite, GigaChat Max and YandexGPT4 models, and to 0.3 for the Mistral NeMo model, as per the developers' recommendations. For the wordplay detection task, the maximum number of generated tokens was set to 128, and for the wordplay interpretation task – to 2,048. For GPT-4o, we used model version *gpt-4o-2024-08-06*, with knowledge up-to-date as of October 2023. The YandexGPT4 model version is specified by its release date, and we used version 23.10.2024. The GigaChat Max version 26.10 was accessed through the API.

For the wordplay detection task, we employed two types of prompts in Russian: 1) a simple prompt asking whether the headline contains wordplay and 2) an extended prompt with definitions and two examples for each of eight wordplay types from the development set, see an extended version of the paper at arXiv.[7] When designing prompts, we adhered to the guidelines outlined in the OpenAI documentation under the *Prompt engineering* section.[8] In both cases, the LLM input included the headline and the lead.

For the wordplay interpretation task, we used 1,033 headlines with annotated *anchor references*, which are not present verbatim in the original headline and thus allow for a streamlined evaluation. The instruction and examples of wordplay were included in the prompt, similarly to the extended prompt in the detection task. In the automatic evaluation, we labeled the interpretation correct if we could match the lemmatized reference in the system's response (the approach is similar to automatic evaluation of pun explanation by Xu et al.).

The results of the experiments are summarized in Table 5. GPT-4o demonstrates the strongest performance in both tasks, significantly outperforming

[7] https://arxiv.org/abs/2503.01510

[8] https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

Model	Detection, P / R		Interpretation, R	
	simple	extended	manual	auto
Giga Lite	0.50 / 0.50	0.53 / 0.72	0.11	0.19
Giga Max	0.62 / 0.48	0.68 / 0.59	0.28	0.28
YaGPT4	0.83 / 0.10	0.76 / 0.24	0.20	0.22
Mistral	0.00 / 0.00	0.00 / 0.00	0.24	0.30
GPT-4o	0.62 / 0.81	0.65 / 0.88	0.48	0.43

Table 5: Wordplay detection precision and recall using a simple/extended prompt and interpretation recall on headlines with anchor references based on manual/string matching scoring.

the other four models. In the detection task, the extended prompt improves both precision and recall in three out of five models. The high precision of YandexGPT's detection comes at the cost of low recall. Interestingly, Mistral returns only noes in the detection task, while it is quite competitive in the interpretation task. YandexGPT4 and GigaChat Max appeared to be very strictly moderated: in the detection task with a simple prompt, they refused to give an answer and suggested changing the topic in 24.8% and 15.4% of cases, respectively.[9]

Looking at the recall of wordplay recognition by type (Table 3), we cannot conclude that some mechanisms are more challenging for all LLMs, but there are small variations within the results of each LLM. The extended prompt improves recall in almost all cases, sometimes significantly. However, the quality of GPT-4o's recognition of oxymorons and polysemy-based wordplay deteriorates slightly. The extended prompt doesn't change GigaChat Max's recall of polysemy-based wordplay. GPT-4o's recall of idiom-based wordplay and nonce words reaches 0.96 – the best among all types. It can be assumed that non-dictionary occasionalisms are very different from the rest of the vocabulary, and idiom-based wordplay is easily recognized due to the frequency and stability of idiomatic expressions it refers to.

Interpretation scores are expectedly lower than detection scores. Again, idiom-based wordplay and nonce words seem to be slightly easier for interpretation; see Table 4. Although not perfect, automatic evaluation seems to be a viable and efficient option in the interpretation task. Lower manual scores are largely due to hallucinations – the models often generate invented phrases that resemble the correct ones.[10] In general, automatic evaluation inflates the scores compared to manual checking, but sometimes the opposite occurs, most notably in case of GPT-4o. We carefully examined these cases and found that GPT-4o returns spelling variants or references that are slightly different from the canonical ones. These cannot be captured by straightforward string matching, but are considered correct by manual evaluation.

The obtained results for both tasks are much lower than LLMs' recognition and explanation scores on English puns (Xu et al., 2024), though they cannot be directly compared.

5 Conclusion

In this paper we presented KoWIT-24, a dataset of richly annotated wordplay in Russian news headlines. We demonstrated how the dataset can be used for wordplay detection and interpretation tasks. The provided multi-level annotation not only contributes to detailed linguistic analysis, but also enables automatic evaluation, which is a significant advantage for NLG tasks. Experiments with five models, which well reflect the variety of available LLMs, show that even advanced models such as GPT-4o face significant challenges in fully understanding and interpreting wordplay in Russian. We expect that the dataset can be used for other tasks as well. For example, previous studies suggest that rich annotation of jokes can improve humor generation (Zhang et al., 2020; Sun et al., 2022; Xu et al., 2024).

We have made the dataset, evaluation scripts, and all code to reproduce the experiments available. We hope that KoWIT-24 will facilitate research in the field of multilingual computational humor.

Acknowledgments. We thank Vladislav Kirienko who developed first version of the script for collecting Kommersant data. PB is partially supported by the Ural Mathematical Center (project #075-02-2025-1719/1). The experiments were partially conducted on HPC facilities at HSE University. Access to YandexGPT4 was possible through Yandex Cloud grant. We thank anonymous reviewers for their insightful comments.

[9]The rejection rate is even higher for more straightforward RIA Novosti headlines – 34.4% and 27.4%, suggesting that Aesopian language can partially overcome strict moderation.

[10]Similarly, when tasked with interpreting RIA Novosti headlines that actually contain no wordplay, LLMs often identify a polysemous word in the text and base their explanation on it.

References

Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. 2023. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13701–13715, Singapore. Association for Computational Linguistics.

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*, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

Britta C. Brugman, Christian Burgers, Camiel J. Beukeboom, and Elly A. Konijn. 2023. Humor in satirical news headlines: Analyzing humor form and content, and their relations with audience engagement. *Mass Communication and Society*, 26(6):963–990.

Davide Buscaldi and Paolo Rosso. 2007. Some experiments in humour recognition using the italian wikiquote collection. In *Applications of Fuzzy Sets Theory*, pages 464–468. Springer Berlin Heidelberg.

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*, pages 7–11, Melbourne, Australia. Association for Computational Linguistics.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Tatyana Chernyshova. 2021. Language mechanisms of building the ironic texts and ways of their linguistic research (linguistic pragmatic aspect). *The European Journal of Humour Research*, 9(1):57–73.

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 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2796–2806. ACM.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part II*, volume 11438 of *Lecture Notes in Computer Science*, pages 87–93. Springer.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> 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, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020b. Stimulating creativity with FunLines: A case study of humor generation in headlines. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 256–262, Online. Association for Computational Linguistics.

Marcio Lima Inacio, Gabriela Wick-Pedro, Renata Ramisch, Luís Espírito Santo, Xiomara S. Q. Chacon, Roney Santos, Rogério Sousa, Rafael Anchiêta, and Hugo Goncalo Oliveira. 2024. Puntuguese: A corpus of puns in Portuguese with micro-edits. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13332–13343, Torino, Italia. ELRA and ICCL.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*.

Pavel Khazanov. 2023. A petrovich inside of every new russian: The disciplinary regime of the capitalist "vanguard group" at 1990s kommersant. *The Russian Review*, 82(3):470–485.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

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 (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

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*, pages

531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tristan Miller, Christian 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 (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Roya Monsefi and Tengku Sepora. 2016. Wordplay in english online news headlines. *Advances in Language and Literary Studies*, 7(2):68–75.

Silviu Oprea and Walid Magdy. 2020. isarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289.

Alan Scott Partington. 2009. A linguistic account of wordplay: The lexical grammar of punning. *Journal of Pragmatics*, 41(9):1794–1809.

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 (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

SaluteDevices team. 2023. rugpt-3.5 13b. technical report. `https://habr.com/ru/companies/sberbank/articles/730108/`, Accessed 15.09.2024.

Hyunju Shin, Isabella Bunosso, and Lindsay R Levine. 2023. The influence of chatbot humour on consumer evaluations of services. *International Journal of Consumer Studies*, 47(2):545–562.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. ExPUNations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4590–4605, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonard Tang, Alexander Cai, and Jason Wang. 2023. The naughtyformer: A transformer understands and moderates adult humor (student abstract). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16348–16349. AAAI Press.

Alexey Tymbay. 2024. Reading 'between the lines': How implicit language helps liberal media survive in authoritarian regimes. the kommersant telegram posts case study. *Discourse & Communication*, 18(4):557–591.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Robert West and Eric Horvitz. 2019. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances"'. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019*, pages 7265–7272. AAAI Press.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Dongyu Zhang, Heting Zhang, Xikai Liu, Hongfei Lin, and Feng Xia. 2019. Telling the whole story: A manually annotated Chinese dataset for the analysis of humor in jokes. 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)*, pages 6402–6407, Hong Kong, China. Association for Computational Linguistics.

Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. 2020. Let's be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*.