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

Anonymous ACL submission

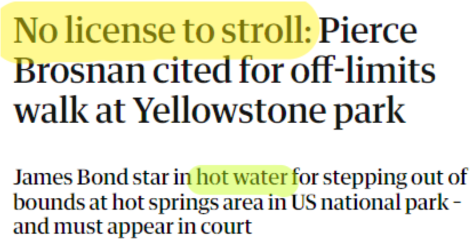
Abstract

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 words/phrases the wordplay refers to. Unlike the majority of existing humor 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://anonymous.4open.science/r/paper-2025-anonymous-submission-65BA/.

1 Introduction

Wordplay refers to creative language use that 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. Humor is a challenging domain for language understanding and generation capabilities of modern LLMs. At the same time, sense of humor is quite applicable: the ability to conduct witty dialogues is a desirable trait of conversational agents (Shin et al., 2023).

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.



No license to stroll: Pierce Brosnan cited for off-limits walk at Yellowstone park

James Bond star in hot water for stepping out of bounds at hot springs area in US national park - and must appear in court

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

Each wordplay-bearing headline is assigned up to two wordplay mechanisms from a set of eight and has annotated *anchor* (wordplay-triggering word or phrase). In addition, we provide a reference 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. 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-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 (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). 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.

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 KoWit-24 Dataset

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 Data Annotation

At the base of KoWit-24 is the binary annotation of the wordplay presence. For the headlines with wordplay, we provide further annotations: 1) wordplay type (each headline can have up to two types), 2) *anchors*, i.e. words or phrases that trigger the wordplay, 3) *anchor reference*, e.g. a similarly 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 Label Studio tool[3] by three authors of the paper, two of whom are professional linguists and one is a computer scientist; all three have an extensive experience with NLP-related projects. An example of the structure of the final dataset is shown in Figure 3 in Appendix D.

For the first phase of annotation, we adopted working definitions of wordplay from the studies of Partington (2009), Attardo (2018), and Laviosa (2015). Besides its *unusuality*, the wordplay should allow for alternative meanings of a text. In this aspect our approach differs, for example, from the works of Monsefi and Sepora (2016), Brugman et al. (2023), where linguistic devices such as personification, metaphor, metonymy, etc. in news headlines are attributed to wordplay. Three annotators labeled 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.

Later, we assigned up to two mechanisms to 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

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

[2] https://www.kommersant.ru/RSS/news.xml

[3] https://labelstud.io/

	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 1: Wordplay types, average anchor length in words (AAL), 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 phrasemes. 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.

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, examples can be found in Appendix A, Table 3.

Finally, we annotated wordplay anchors, provided anchor references, and, if possible, added a link to the corresponding Wiktionary or Wikipedia page.[4] 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. The overlapping annotation was particularly useful in this stage: the different cultural preferences and backgrounds of the annotators allowed to get a higher coverage, as not all references are obvious and immediately understandable.

3.3 Dataset Statistics and Analysis

In total, we annotated 2,700 headlines, of which 1,340 contained wordplay, so the dataset is almost perfectly balanced. 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 $\kappa = 0.58$). However, we hope that the implemented procedure (triple overlap and subsequent reconciliation of discrepancies) ensures a high quality of the resulting annotation. It is interesting to note that the wordplay headlines are on average one word shorter than their counterparts (3.88 vs. 4.81 words). We calculated perplexity of the headlines in both KoWit-24 classes and 1k headlines from the RIA Novosti dataset (Gavrilov et al., 2019) using ruGPT-3.5 13B model (SaluteDevices team, 2023). 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.

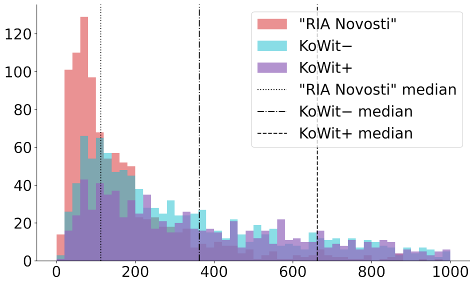


Figure 2: Perplexity distribution of the headlines in two (+/–) KoWit-24 classes and RIA Novosti collection. Vertical lines correspond to the medians of the distributions (note that the histogram is truncated).

Distribution of headlines by wordplay type can be seen in Table 1. The most frequent wordplay mechanism in our dataset appeared to be the modification of existing well-known phrases – collocations, idiomatic expressions, or named entities. The second meaning, a necessary condition for wordplay, arises not from the polysemy or phonological similarity as in puns, but as a reference to the source phrase. 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 proposed a conceptual framework for describing the structure and function of wordplay as occurring mainly at the *phrasal* level and applied this framework to the analysis of wordplay in a large collection of British news headlines.

About a 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.

4 Experiments

For the experiments, we allocated 200 records (100 from each class) for the development set, mak-

[4] 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 for the first entry.

ing sure that all wordplay types were represented. Thus, the test set contains 2,500 headlines (1,290 with and 1,310 without wordplay).

We experimented with two tasks – 1) wordplay detection and 2) wordplay interpretation. We employed five LLMs: GPT-4o, Mistral NeMo 12B, YandexGPT4, GigaChat Lite, and GigaChat Max. These five are a representative mix of open/closed, medium-sized/large, and Russian-centric/multilingual models. Details about the LLMs can be found in Appendix B.

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 definition and two examples for each of eight wordplay types from the development set, see Table 5 in Appendix B. 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 allows 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 2. GPT-4o demonstrates the strongest performance in both tasks, significantly outperforming the other four models. In the detection task, the extended prompt improves both precision and recall in three out of five models (see detailed results in Table 6 in Appendix C). 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 outperforms both GigaChat versions and YandexGPT4 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.[5]

Although not perfect, automatic evaluation seems to be a viable and efficient option in the interpretation task, see detailed results in Appendix C.

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 2: 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 difference with manual evaluation is largely due to hallucination – the models often generate invented phrases that resemble the correct ones. Notably, manual evaluation increases GPT-4o's scores, while decreasing the scores of the other models, which is to be expected. We carefully examined these cases and found that OpenAI's models return spelling variants or references that are slightly different from the canonical ones, but are considered correct.

The obtained results for both tasks are much lower than LLLs' 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.[6] We hope that KoWit-24 will facilitate research in the field of multilingual computational humor.

[5] 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.

[6] https://anonymous.4open.science/r/paper-2025-anonymous-submission-65BA/

6 Limitations

There are several limitations to study wordplay in headlines. First, the annotation process is inherently subjective, as the identification of wordplay may vary depending on individual interpretation, educational background, etc. However, we hope that the implemented procedure ensures a high quality of the resulting annotation. Second, the specific editorial style of Kommersant introduces bias, as the outlet is known for its particular style and language, which may not be representative of broader journalistic practices. In addition, the experiments used only five LLMs and did not involve extensive prompt engineering, meaning that the reported results can potentially be improved with more effective prompts and the use of different LLMs.

7 Ethical considerations

Our dataset reveals instances of wordplay even in the headlines of articles about sensitive topics such as diseases, death, and war, that some readers may find unacceptable. We will add a warning to the published dataset.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Salvatore Attardo. 2018. Universals in puns and humorous wordplay. In Esme Winter-Froemel and Verena Thaler, editors, *Cultures and traditions of wordplay and wordplay research*, pages 89–110. De Gruyter Berlin & Boston.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

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.

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.

Sara Laviosa. 2015. Wordplay in advertising: Form, meaning and function. *Scripta Manent*, 1(1):25–34.

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.

Mistral AI team. 2024. Mistral nemo. `https://mistral.ai/news/mistral-nemo/`, Accessed 15.09.2024.

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

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.

SaluteDevices team. 2024a. Gigachat lite. technical report. `https://habr.com/ru/companies/sberdevices/articles/865996/`, Accessed 09.02.2025.

SaluteDevices team. 2024b. Gigachat max. technical report. `https://habr.com/ru/companies/sberdevices/articles/855368/`, Accessed 09.02.2025.

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.

Yandex team. 2024. Yandexgpt 4. https://ya.ru/ai/gpt-4, Accessed 09.02.2025.

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*.

A Wordplay Examples

Wordplay type	Original headline	Transliteration
Polysemy	«Волгу» не могут заставить течь быстрее	Volgu ne mogut zastavit' tech' bystree
Homonymy	Туризм подрастерял Шарм	Turizm podrasteryal Sharm
Homophony	Из-под земли до стали	Iz-pod zemli do stali
Collocation	Особо бумажные персоны	Osobo bumazhnye persony
Idiom	Код накликал	Kod naklikal
Reference	Миссия сократима	Missiya sokratima
Nonce word	От запчастного к общему	Ot zapchastnogo k obshchemu
Oxymoron	Новый премьер Израиля начал со старого	Noviy premier Izrailya nachal so starogo

Table 3: Wordplay mechanisms with original examples and their transliterations.

B LLM Usage Details

Model	Availability	xMMLU,%	Release date	N
GigaChat Lite (SaluteDevices team, 2024a)	Open	58.38^{α}	2024-12-13	20B
GigaChat Max (SaluteDevices team, 2024b)	Closed	75.00^{α}	2024-11-02	?
YandexGPT4 (Yandex team, 2024)	Closed	65.00^{β}	2024-10-23	?
Mistral Nemo (Mistral AI team, 2024)	Open	59.20^{α}	2024-07-18	12B
GPT-4o (Achiam et al., 2023)	Closed	86.40^{γ}	2024-08-06	?

Table 4: Five LLMs in the study. xMMLU scores refer to various versions of the original MMLU benchmark: α – ruMMLU[7] is an open Russian version, β – yaMMLU, a proprietary Russian version by Yandex (Yandex team, 2024), γ – original English MMLU (Hendrycks et al., 2021); N – total number of model parameters.

We selected YandexGPT, GigaChat Lite, and GigaChat Max as they are specifically optimized for Russian language processing. Among these, GigaChat Lite[8] is open-source, with model weights distributed under the MIT license. Additionally, we included Mistral Nemo[9], an open model that provides official support for the Russian language and is distributed under the Apache 2.0 license. Finally, we incorporated GPT-4o due to its status as one of the leading models in the industry. Additional details about the models are provided in Table 4.

When using LLMs, we set the following hyperparameters: temperature and the maximum number of generated tokens. 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, it was set 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 employed through the API.

Table 5 shows examples of the prompts used. Full text of prompts and instruction can be found in the github repository.[10] Note that the simple prompt matches the head of the extended one (before the <instruction> part with definitions and examples).

[7] https://github.com/NLP-Core-Team/mmlu_ru

[8] https://huggingface.co/ai-sage/GigaChat-20B-A3B-instruct-v1.5

[9] https://huggingface.co/mistralai/Mistral-Nemo-Base-2407

[10] https://anonymous.4open.science/r/paper-2025-anonymous-submission-65BA/

Prompt type	**Original prompt**	**Translated prompt**
User prompt	Заголовок новости: <headline>. Содержание новости: <lead>	Headline: <headline>. News content: <lead>
System prompt for wordplay detection	Присутствует ли в заголовке новости игра слов? Дай ответ с учетом содержания новости. Отвечать можешь только 'да', 'нет' или 'не знаю'. <instruction>	Does the news headline contain wordplay? Give an answer considering the content of the news. You can only answer 'yes', 'no' or 'don't know'. <instruction>
System prompt for wordplay interpretation	Проанализируй заголовок новости в контексте ее содержания. Укажи, есть ли в заголовке игра слов. Если она есть, объясни смысл, использованные методы и связь с основным текстом. Если игры слов нет, то ответь "в заголовке нет игры слов". <instruction>	Analyse the news headline in the context of its content. Identify whether there is wordplay in the headline. If there is, explain the meaning, the methods used and the relationship to the main text. If there is no wordplay, answer 'there is no wordplay in the headline'. <instruction>

Table 5: User and system prompts for wordplay detection and interpretation.

C Wordplay Detection and Interpretation Results by Wordplay Type

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

Table 6: 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*		*GPT-4o*		*Mistral*	
		manual	auto	manual	auto	manual	auto	manual	auto	manual	auto
Polysemy	12	0.17	0.33	0.33	0.25	0.08	0.00	**0.50**	0.50	0.17	0.33
Homonymy	7	0.14	0.43	0.29	0.29	0.14	0.14	**0.43**	0.29	0.00	0.43
Phonetic similarity	85	0.11	0.25	0.28	0.39	0.11	0.21	**0.52**	0.51	0.15	0.32
Collocation	393	0.07	0.18	0.27	0.27	0.16	0.20	**0.44**	0.41	0.20	0.30
Idiom	164	0.15	0.18	0.37	0.30	0.32	0.28	**0.55**	0.48	0.34	0.30
Reference	326	0.10	0.12	0.25	0.23	0.20	0.16	**0.46**	0.36	0.23	0.25
Nonce word	166	0.15	0.46	0.29	0.57	0.28	0.43	**0.61**	0.69	0.28	0.57
Oxymoron	6	0.67	0.67	0.50	0.50	0.33	0.33	0.67	0.50	**0.83**	0.83

Table 7: Recall on wordplay interpretation by type; manual and automatic evaluation.

D Dataset Structure Example

```
headline: Диалектический пиломатериализм
lead: Цены на фанеру и доски начали снижаться вслед за спросом
summary: Пиломатериалы и лесопромышленная продукция начинают дешеветь по
    мере завершения строительного сезона. По мнению аналитиков и некоторых
    участников рынка, этому способствует сокращение спроса на фоне летнего
    всплеска цен. И хотя на некоторые продукты, например OSB, цена упала уже
    на треть, она все еще вдвое выше уровня конца прошлого года. До конца года
    можно ожидать стабилизации цен, полагают участники рынка, но едва ли
    возвращения к средним многолетним значениям.
is_wordplay: yes
date: 2021-10-27
article_url: https://www.kommersant.ru/doc/5051268
annotations: [
        {
            headline_substring: Диалектический пиломатериализм
            start_index: 0
            end_index: 30
            wordplay_type: reference
            reference_string: Диалектический материализм
            reference_url: https://ru.wikipedia.org/wiki/Диалектический_материализм
        },
        {
            headline_substring: пиломатериализм,
            start_index: 15
            end_index: 30
            wordplay_type: nonce word
            reference_string: [
                материализм,
                пиломатериалы
            ]
        }
]
```

Figure 3: Dataset entry example (based on original JSON format, simplified for readability).