


UDC 81'322

DOI: 10.18413/2313-8912-2024-10-4-0-4

Elena S. Rudenko<sup>1</sup> 

Marina Yu. Semenova<sup>2</sup> 

**Artificial vs human intelligence:  
a case study of translating jokes based on wordplay**

<sup>1</sup> Don State Technical University,  
1 Gagarina Sq., Rostov-on-Don, 344000, Russia  
E-mail: [spu-47.5@donstu.ru](mailto:spu-47.5@donstu.ru)  
ORCID: 0000-0003-3552-4034

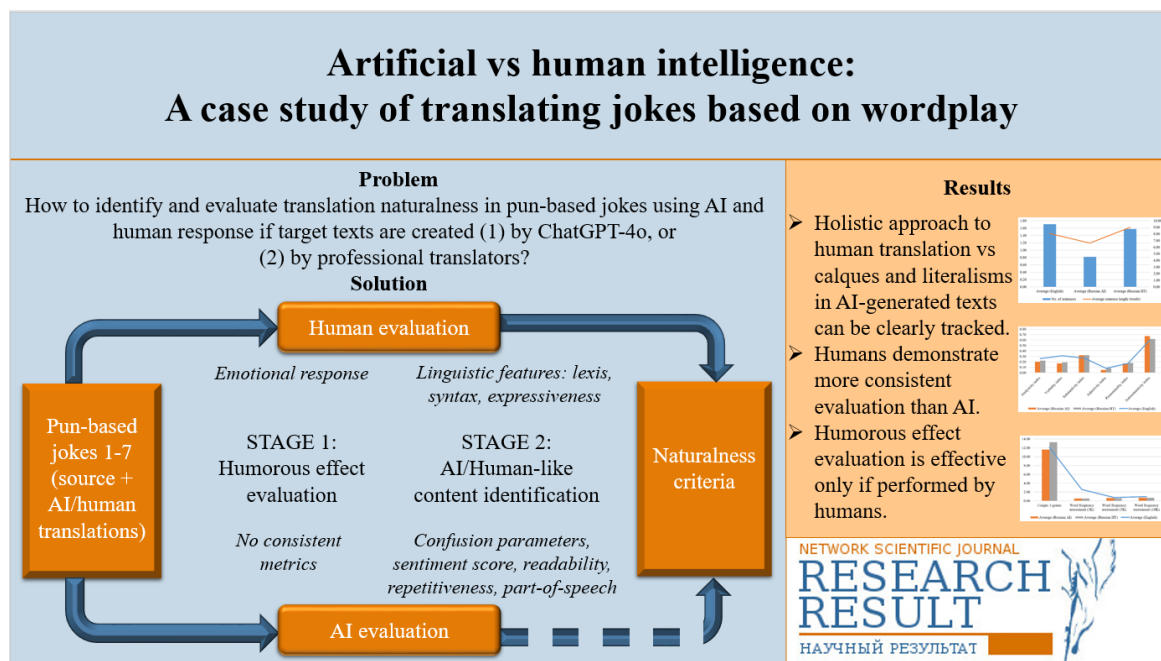
<sup>2</sup> Don State Technical University,  
1 Gagarina Sq., Rostov-on-Don, 344000, Russia  
E-mail: [spu-47.5@donstu.ru](mailto:spu-47.5@donstu.ru)  
ORCID: 0000-0003-2654-2409

*Received 01 September 2024; accepted 15 December 2024; published 30 December 2024*

**Abstract:** Artificial intelligence (AI) technologies used in professional translation question the effectiveness of human-AI interaction. Deep learning can mimic human cognitive processes, accordingly suggesting that AI could reproduce the logic and mechanics of the source text in the target language. The study necessitates an objective assessment of the neural machine translation (NMT) naturalness, which will apply prompt engineering to optimize the translation process, save resources, and ensure the sustainable development of super-central and central natural languages of the world. The study employs English rhyming/non-rhyming pun-based jokes, and the corresponding Russian translations performed by both professional translators and by ChatGPT-4o, with the prompts for human and AI translators being the same. The results obtained were processed using linguistic and translation analysis followed by textometric and statistical analysis. To evaluate the humorous effect of the translated jokes and to identify signs of artificiality in these jokes, 150 informants were surveyed. The study established the degree of humorous effect and the naturalness criteria for the translated jokes. While the source text lacks terminology, specialized words and complex grammar, the AI-generated translations were perceived as complex due to literalisms and calques. Conversely, human translators prefer a holistic translation technique and are more flexible to interpret imagery and syntactic structures of jokes. This highlights a greater creative freedom of human translators, who avoid stereotypes and generate novel interpretations. In conclusion, the study measures the effectiveness of AI as an auxiliary tool for translating and assessing pun-based jokes.

**Keywords:** Artificial intelligence; Prompt engineering; Naturalness of translation; Wordplay; Pun-based jokes; Neural machine translation efficiency

**How to cite:** Rudenko, E. S., Semenova, M. Yu. (2024). Artificial vs Human Intelligence: A Case Study of Translating Jokes Based on Wordplay, *Research Result. Theoretical and Applied Linguistics*, 10 (4), 53-93. DOI: 10.18413/2313-8912-2024-10-4-0-4



УДК 81'322

DOI: 10.18413/2313-8912-2024-10-4-0-4

Руденко Е. С.<sup>1</sup>   
Семенова М. Ю.<sup>2</sup>

Искусственный интеллект против естественного:  
кейс-стади перевода шуток, основанных на игре слов

<sup>1</sup> Донской государственный технический университет  
пл. Гагарина, 1, Ростов-на-Дону, 344000, Россия  
E-mail: [spu-47.5@donstu.ru](mailto:spu-47.5@donstu.ru)  
ORCID: 0000-0003-3552-4034

<sup>2</sup> Донской государственный технический университет  
пл. Гагарина, 1, Ростов-на-Дону, 344000, Россия  
E-mail: [spu-47.5@donstu.ru](mailto:spu-47.5@donstu.ru)  
ORCID: 0000-0003-2654-2409

Статья поступила 10 сентября 2024 г.; принята 15 декабря 2024 г.;  
опубликована 30 декабря 2024 г.

**Аннотация:** Внедрение технологий искусственного интеллекта в профессиональный перевод ставит вопрос об эффективности взаимодействия искусственного и естественного интеллекта. Глубокое обучение способно имитировать когнитивные процессы, присущие человеку. Соответственно, можно предположить, что искусственный интеллект способен воспроизвести логику и механику исходного текста на переводящем языке. Актуальность исследования обусловлена необходимостью объективно оценить степень естественности нейронного машинного перевода. Подобный подход позволит с помощью промпт-инжиниринга оптимизировать процесс перевода, сэкономить ресурсы и обеспечить устойчивое развитие естественных языков, занимающих центральные и суперцентральные позиции в общей иерархии языков мира. Исследование проведено на материале англоязычных рифмованных и

нерифмованных шуток, основанных на игре слов, и переводе данных шуток на русский язык, выполненном как профессиональными переводчиками, так и нейросетью ChatGPT-4o. При этом важно, что при выполнении поставленной задачи промпты для естественного и искусственного интеллекта совпадали. Полученные результаты прошли обработку методами лингвистического и переводческого анализа с последующей текстометрической и статистической обработкой. Для определения степени выраженности юмористического эффекта переведенных шуток и для идентификации признаков искусственности в текстах шуток было опрошено 150 информантов. В результате была установлена степень выраженности юмористического эффекта в переведенных шутках и определены критерии естественности перевода. На фоне отсутствия терминологии, специальных слов и сложных грамматических структур в тексте на языке-источнике, искусственно сгенерированный текст перевода воспринимался информантами как сложный за счет буквализмов и калек. Естественные тексты перевода основаны на целостном преобразовании и характеризуются гибкостью интерпретации образов, подвижностью синтаксических структур. Это свидетельствует о большей творческой смелости переводчика-человека, способного уйти от стереотипов и создать новый, незаштампованный образ. В заключение оценивается эффективность искусственного интеллекта как вспомогательного инструмента при переводе и оценке шуток, основанных на игре слов.

**Ключевые слова:** Искусственный интеллект; Промпт-инжиниринг; Естественность перевода; Шутки, основанные на игре слов; Эффективность нейронного машинного перевода

**Информация для цитирования:** Руденко Е. С., Семенова М. Ю. Искусственный интеллект против естественного: кейс-стади перевода шуток, основанных на игре слов // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 4. С. 53–93. DOI: 10.18413/2313-8912-2024-10-4-0-4



## 1. Introduction

Translation has long become one of the major spheres of AI application. First attempts to employ machines to convey the logic and mechanics of the source text in the target language date back to the pre-digital era of the first half of the 20<sup>th</sup> century (automatic translation term base system invented and patented by P. Smirnov-Troyansky, 1933). First digital approaches to tackling text generation for translation purposes emerged in the 1950s as rule-based machine translation, which later evolved into example-based machine translation in the 1980s. The 1990s saw a new approach to machine translation based on statistics which dominated the sphere well through the 2000s (first Google Translate and similar solutions) up to 2015 when the neural network approach took the leading part. Since 2016, General Pre-Trained Transformer (GPT) has been introduced not only in amateur translation, but also has become an indispensable tool for professional translators.

The past decade has formulated a number of interdisciplinary issues concerning human-AI interaction, e. g.: AI ethics, text generation authorship rights, translation quality control management, etc. At the same time, specific psycholinguistic issues instigate research into human vs AI cognitive processes, text evaluation, and AI detection, which, in turn, requires new assessment tools for the neural machine translation (NMT) naturalness. Prompt engineering has proven effective for translation process optimization as it saves resources and stimulates linguistic research of languages across the globe with a special focus on so-called super-central and central natural languages. Such languages occupy leading positions at the national and international levels. English and Russian represent one of the most popular translation pairs in the world, with both languages competing with each other to occupy the super-central position. This competition covers human communication issues, but also takes place in the AI-related sphere. Today, the question arises whether the status of

high/low resource language is true for all AI-related tasks as there is evidence that even high resource languages might lack reliable solutions to identify AI-generated texts. It is a challenge for the Russian language as well since there is a global need to detect and avoid misinformation and fake news and to achieve greater accuracy of information in texts that we produce and receive. Our awareness related to AI-generated texts has become a pressing issue, as it imposes certain ethical implications in case of blind usage of AI-like content, and relies heavily on the belief that humans are able to evaluate texts for their naturalness.

## 2. Background

To compare AI-generated and human translations, linguists need to tackle several issues.

First of all, there is a lack of knowledge on whether large language models (LLMs) can employ a translation process in the same or similar way as human translators. In academic circles, there is a common understanding that “current LLMs are approaching human-like general intelligence, the extent to which LLM can emulate such strategies remains underexplored” (He et al., 2023). Yet the discussion of whether LLMs’ language ability can be understood in the same terms as human thinking is still open as researchers state that “LLMs have poor reasoning skills despite possessing human-level language skills” (Bang et al., 2023).

Dealing with AI-generated translations requires a deep understanding of prompt engineering approaches. One of the basic features of a prompt is its component structure, including but not limited to the following: directive, output formatting and other instructions, style/genre instructions, role. The focus on translation-related issues requires a deeper insight into roles. According to recent research, there are two design principles: speaking style imitation (lexical consistency and dialogic fidelity) and role-specific knowledge and memory injection (script-based and script-agnostic knowledge)

(Wang et al., 2023). These criteria can provide responses that ensure lexical alignment with the role as well as infuse role-specific knowledge.

Another issue regarding prompt engineering is that of understanding human texts. Undoubtedly, “Generative Pre-trained Transformers (GPTs) have revolutionized natural language processing. Each iteration pushes AI language models forward with transformative capabilities” (Pan et al., 2017). At the same time, it should be mentioned that communication always originates in response to social motivations (Pfaff, 1979). Thus, the question is whether it is plausible to monitor social factors and register social response in human-AI communication and whether these parameters are able to prevent users from functional and syntactic constraints as they might “reflect semantic and communicational properties of discourse” (ibid.).

Among a range of text generation tasks, machine translation traditionally is considered at the sentence and contextual levels and presupposes such challenges as: ambiguities, low-resource languages, and long sentences (Becker et al., 2024).

It is worth to note that another important task here is automatic text evaluation. Today, there are dozens of online tools, although they are reported to be unreliable, providing a 50% quality score, which means that their measure is general and the prediction is random (Pan et al., 2017). It is claimed that “text evaluation aims to assess the quality of hypothesis text  $h$  in terms of certain aspect  $a$  (e. g., fluency), which is either measured manually with different protocols <...> or quantified by diverse automated metrics” (Fu et al., 2023).

One of the main issues concerning the inconsistency of automatic systems is that in many cases they rely on n-gram overlap between two texts, which does not take into account meaning-preserving lexical and compositional diversity (Zhang et al., 2020).

Traditionally, human evaluations are assumed to be the best to identify and assess a text. The obvious drawbacks of human

evaluation are the following: expensiveness, high latency, and inability to fit in a daily model development pipeline (Sellam et al., 2020). In contrast, AI detectors are cheaper and more accessible, which makes them instrumental for quantifying preliminary or final results and for the system optimization (Celikyilmaz et al., 2021; Yang, Wang et al., 2023).

Human or AI evaluation is always a challenging task (Goddard et al., 2024) as the object we deal with here is an open-ended product. It means that the same input can result in multiple responses as an output, so human evaluators are supposed to perform much better due to their flexibility and creative freedom. Fundamentally, this means that automatic metrics just tries to replicate human decisions and cannot be as important and self-sufficient as a human (ibid.). Nevertheless, there is evidence that humans also detect AI-like content at chance level. Some studies dealing with the issue of AI-generated essay identification reveal that human evaluation accuracy is 59% on average compared to 61% for ESL teachers and taking in account the maximum level of 67%, which requires exposure and self-training (Fraser et al., 2024).

Papers dealing with human evaluation issues usually compare translation capabilities of GPT models and humans varying in terms of language pairs and translation directions (Puduppully et al., 2023; Etxaniz, et al., 2023). The problem is that there is no uniform approach to the metrics, e. g.:

- unaligned target/source words, punctuation, monotonicity (Hendy et al., 2023);

- quantitative analysis: part of speech (POS) and sentence length (Yang, 2023; Doughman et al., 2024);

- word frequency, sentence length (Jiao et al., 2023);

- number of whitespaces, empty line, inline whitespaces, punctuations, trailing whitespaces, lines with leading whitespaces, maximum length (Oedingen et al., 2024);

- accuracy, coherence, readability, and human likeliness (Çano, Bojar, 2020);

- syntactic and lexical diversity; repetitiveness; coherence; purpose (Gryka, 2024);

- intrinsic methods assessing similarity of the systems' output to a reference model or quality criteria or employing user like measure (Likert or rating scales); extrinsic methods based on user task success metrics or system purpose success metrics: An evaluation type where a given system is evaluated by measuring whether it can fulfil its initial purpose (Gkatzia, Mahamood, 2015);

- connect/disconnect (the text looks right/wrong) (Schuff et al., 2023), etc.

To distinguish between AI-generated and human-translated texts, we also need to understand the extent to which it is possible to apply differentiation criteria.

According to current academic research analysis, the number of metrics used to create an estimation model can reach 130, including such linguistic characteristics as “lexical, semantic, and syntactic properties of a text, its coherence, as well as sequences of part-of-speech tags, some word-formation patterns, and general-language frequency of lemmas”, etc. (Blinova, Tarasov, 2022).

AI-generated texts produce a new academic issue of naturalness vs artificiality. Before GPT, translation naturalness used to be associated with ‘taste’ and characterized by lack of objectivity in terms of grammar and vocabulary (Rogers, 1999). Today, academic literature features a different understanding of naturalness, which is now described in terms of accuracy, clearness and flow. Another shift in meaning concerns the sphere of application of the naturalness criterion: from students' academic works to literary translations. Therefore, the academic polemics concentrates on the correlation between attraction and naturalness of the text from the reader's viewpoint (Fadaee, 2011). Another modern approach to naturalness correlates it to accuracy and introduces three types of translation errors according “to the tension

between naturalness and accuracy, which are: natural – inaccurate, unnatural – accurate, and unnatural – inaccurate” (Obeidat et al., 2020).

Being an issue of psycholinguistic nature, the appeal of translated texts to the audience should include a wide range of attributes, among which emotional intelligence has already become one of the major cornerstones. It can be assumed that naturalness cannot be achieved without an effective interpretation and management of emotion-infused information (Li et al., 2023).

### 3. Aim of the study

The study is aimed at devising an objective approach to identify and evaluate translation naturalness of pun-based jokes for two sets of translations: neural machine translations (NMT) and those performed by humans. Meeting the naturalness requirement via prompt engineering will facilitate the optimization of language resources and sustainability for English-Russian NMT. Both languages belong to central natural languages currently competing to occupy the position of a super-central natural language of the world. In terms of NMT, it means that the study deals with a high-resource language pair.

To evaluate the naturalness of translation, we need to perform the following tasks:

- 1) to define the humorous effect of the jokes;
- 2) to identify signs of artificiality in the jokes;
- 3) to establish the degree of humorous effect and the naturalness criteria for the translated jokes;
- 4) to measure the effectiveness of NMT as an auxiliary tool for translating pun-based jokes.

### 4. Materials and methods

#### 4.1. Stage 1

This study analyzed seven English rhyming/non-rhyming jokes, which are based on wordplay (pun in particular) and devoid of opaque or complex words and contexts. Consequently, in the collective unconscious

they are not regarded as “intellectual”, i.e., requiring profound exegesis, and are often referred to as ‘dad jokes’. The linguistic mechanisms underlying these jokes are transparent and straightforward, typically relying on ambiguities at the phonological, morphological, or semantic levels. Resolving those ambiguities (e. g., incongruity between literal and figurative meaning of words) triggers a response of laughter. Structurally, pun-based jokes consist of set-up phrase followed by a punch line. The jokes selected for this study were chosen based on two criteria:

1) “clash-of-languages” criterion focusing on systemic differences between English and Russian, which pose a particular difficulty for translators since the form and content of the respective source and target languages do not coincide one-to-one;

2) novelty criterion ensuring that the jokes selected have not been previously translated into Russian.

The list of jokes is presented below:

1. *What did the pirate say when he turned 80? Aye matey!*

2. *Candy is dandy but liquor is quicker.*

3. *I took the shell off my racing snail, thinking it would make him run faster. If anything, it made him more sluggish.*

4. *Two windmills are standing in a field and one asks, “What’s your favorite kind of music?” The other says, “I’m a big metal fan”.*

5. *“I have a split personality”, said Tom, being frank.*

6. *I tried catching fog yesterday. Mist.*

7. *I can’t believe I got fired from the calendar factory. All I did was take a day off.*

Stage 1 applies linguistic analysis to look into the humorous effect created in these pun-based jokes English jokes and to define linguistic features of each of the English jokes.

Linguistic features processed via this analysis which are further statistically aggregated using the textometric approach (Corizzo, Leal-Andreas, 2024):

1) text: type and numbers of punctuation marks, average sentence length;

2) repetitiveness: unique n-gram quantity<sup>1</sup>, word frequency assessment normalized to 1 (comparison to the 3K, 5K and 10K most used words in the given language)<sup>2</sup>;

3) emotional semantics: sentiment score<sup>3</sup>;

4) readability<sup>4</sup>: automated readability index; reading difficulty; grade level; age range; Flesch-Kincaid index; Coleman-Liau index;

5) part-of-speech (POS): frequency of certain word types in the text (analyticity index, verbality index, substantivity index, adjectivity index, pronominality index, autosemanticity index).

The dataset as well as linguistic analysis data for the presented research is available online<sup>5</sup>.

#### 4.2. Stage 2

Stage 2 involves the translation process followed by the translation analysis and performed for the resulting jokes in Russian which were generated by human and AI translators separately.

Human translations (HT) were carried out by fourth-year students enrolled in the “Translation and Translation studies” educational program at Don State Technical University (Rostov-on-Don, Russia), in liaison with the corresponding teaching staff providing expertise; the translations were produced as part of the “Translation Analysis”

<sup>1</sup> Unique n-grams frequency counter: <https://corpus.by/NgramFrequencyCounter/>

<sup>2</sup> English word lists: <https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000/>; <https://word-by-word.ru/ratings/top-10000-words>

Russian word lists: [https://ru.wiktionary.org/wiki/Приложение:Список\\_частотности\\_по\\_НКРЯ](https://ru.wiktionary.org/wiki/Приложение:Список_частотности_по_НКРЯ)

<sup>3</sup> Sentiment scores: <https://text2data.com/Demo/>; <https://products.groupdocs.app/classification/ru/text>

<sup>4</sup> Readability scores: <https://readabilityformulas.com/>; <https://readability.io>

<sup>5</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation>

course. Students were instructed regarding procedures to be followed in translating jokes involving pre-drafting, drafting and post-drafting (Mossop, 2000).

To objectively assess the effectiveness of neural machine translation (NMT) as an auxiliary tool for translating pun-based jokes, we utilized the latest ChatGPT-4o model.

To elicit the desired responses – translations produced by ChatGPT-4o – we used a prompting template called cross-lingual thought prompting (XLT), introduced by Wei et al. and aimed at stimulating “cross-lingual and logical reasoning skills to enhance task performance across languages.” (Wei et al., 2022) The principal criterion for choosing this particular method was the fact that it is “effective under multilingual scenarios” (Shi et al., 2023). The template is comprised of the following six logical instructions that guide the model’s behavior:

*I want you to act as a task\_name expert for task\_language.*

*task\_input*

*You should retell/repeat the input\_tag in English.*

*You should task\_goal.*

*You should step-by-step answer the request.*

*You should tell me the output\_type (output\_constraint) in this format 'output\_type:' (ibid.).*

Based on this template, we asked ChatGPT-4o to think step-by-step to solve a translation task, an example of instantiated prompt is given below:

*I want you to act as a translation expert for Russian.*

*Request: В данном диалоге я хочу, чтобы ты действовал как опытный переводчик с английского языка на русский. Твоя задача: сохранить при переводе шутки с английского языка на русский игру слов, т.е. адекватный юмористический эффект. Если игру слов на английском языке невозможно передать на русский язык через идентичный образ, подбери новый эквивалент на других основаниях, который позволил бы сохранить игру слов.*

*Тебе нужно перевести следующую шутку с английского языка на русский: 'What did the pirate say when he turned 80? Aye matey!'*

*You should retell the request in English.*

*You should do step-by-step joke analysis to translate the joke from English into Russian.*

*You should step-by-step answer the request.*

*You should tell me the joke translation in this format 'Перевод шутки:'.*

In most cases, the initial translations generated by ChatGPT-4o did not fulfill the communicative purpose of the joke. As a result, further dialogue was needed to give refinement to translations, ensuring that the jokes retained their humorous effect in Russian; the contextual dialogue involved different conversation techniques:

- expressing disagreement with the results obtained and asking the model to generate two or three more translations;

- explaining the reason why the joke translation was not amusing and asking the model to retranslate the joke with regard to explanations;

- asking follow-up questions;

- encouraging GPT to improve its effectiveness etc.

The dialogues with ChatGPT-4o varied in terms of length and techniques used, continuing until a pronounced satisfactory humorous effect was achieved. The instantiated dialogues are provided in Appendixes 1-8<sup>6</sup>.

### 4.3. Stage 3

Stage 3 focuses on linguistic features and corresponding confusion parameters of each of the Russian jokes generated by human and AI translators, which characterize the humorous effect created in Russian. It is important to note that these are the same linguistic features and the same approaches as those employed at Stage 1 for the English dataset.

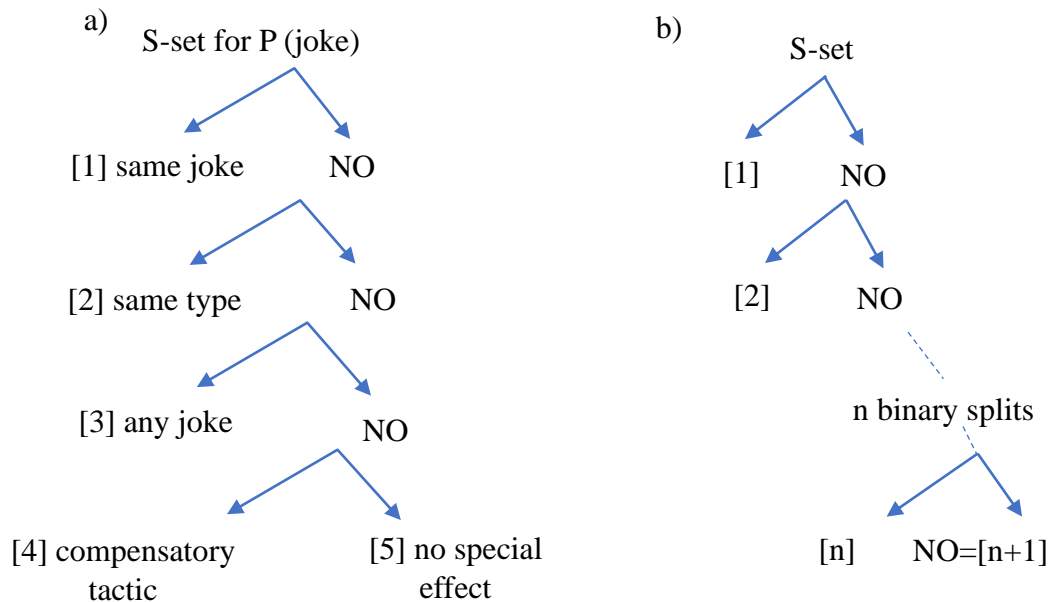
<sup>6</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/tree/main/Appendixes>



To assess the degree of humorous effect in translated jokes, we employed a binary branching model for structuring joke-types

(Zabalbeascoa, 2005), which enables us to classify a set of translations generated by LLM into jokes and non-jokes (see Figure 1).

**Figure 1.** Set of solutions S: binary branching tree structure for translating problem P.  
**Рисунок 1.** Бинарное дерево переводческих решений S для переводческой задачи P.



It should be noted that some jokes' translation solutions are dependent on a particular translation practice when it is important for a translator to understand what constitutes the same joke, same type of humour etc. This requires to adapt the following classification of translation solution types to suit our objectives concerning the translation and evaluation of jokes:

- 1) maintaining the initial play element and/or imagery;
- 2) partial loss of the initial play element and/or imagery;
- 3) using a different play element and/or imagery;
- 4) loss of the initial play element, resorting to compensatory tactics – using other stylistic or structural means (similes, hyperbolae etc.);
- 5) complete loss of the initial play element without compensation.

Translations, produced by ChatGPT-4o, were evaluated as adequate if they fell into solution-types 1-4.

Human and automatic evaluations were processed using the following approach to statistically evaluate naturalness/artificiality criteria applied in this paper (adopted to the study from (Pan et al., 2017)):

1. True Positive Rate (TPR/Recall) –  $TPR/Recall = \frac{TP}{TP+FN}$ , where:

TP = number of human-translated jokes correctly labelled as human-translated,

FN = number of human-translated jokes incorrectly labelled as AI-translated,

TP+FN = the total number of human-translated jokes

2. False Negative Rate (FNR) –  $FNR = \frac{FN}{TP+FN}$ , where:

FN = number of human-translated jokes incorrectly labelled as AI-translated,

TP = number of human-translated jokes correctly labelled as human-translated,

TP+FN = the total number of human-translated jokes.

3. True Negative Rate (TNR) –  $TNR = \frac{TN}{TN+FP}$ , where:

TN = number of AI-generated translations correctly labelled as AI-translated,

FP = number of AI-generated translations incorrectly labelled as human-translated.

TN+FP = the total number of AI-generated translations.

4. False Positive Rate (FPR) –  $FPR = \frac{FP}{TN+FP}$ , where:

FP = number of AI-translated jokes incorrectly labelled as human-translated,

TN is the number of AI-generated translations correctly labelled as AI-translated,

TN+FP = the total number of AI-generated translations.

5. Accuracy (ACC) –  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ , where:

TP = number of human-translated jokes correctly labelled as human-translated,

TN = number of AI-generated translations correctly labelled as AI-translated,

TP+TN+FP+FN = the total number of jokes.

6. Precision –  $Precision = \frac{TP}{TP+FP}$ , where:

TP = number of human-translated jokes correctly labelled as human-translated,

FP = number of AI-generated translations incorrectly labelled as human-translated,

TP+FP = the total number of jokes labelled as human-translated.

7. F1 Score –  $F1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$ , which represents the harmonic mean of precision and recall.

This stage also identifies signs of artificiality in the Russian jokes via global translation strategy analysis as well as measures the effectiveness of NMT relying on human and automatic evaluations.

The humorous effect and naturalness of both human- and AI-powered translations was measured using questionnaire, consisting of a set of questions given below.

- Are you male or female?

- How old are you?

- Please rate your command of English on a scale of 1 to 10 (1 being lowest, and 10 being highest).

- Which joke do you think is the funniest?

- Which joke translation(s) do you believe were generated by AI?

- Which words or expressions in the translation(s) of joke (number) indicate that the text might have been generated by AI?

Additional remarks should be made regarding the design of the questions.

The question “Which joke do you think is the funniest?” was presented in a multiple-choice format offering 3-5 translation variants produced by both human and AI-translators intermixed. An important point we would like to emphasize is that the respondents in this section of the survey were not informed that the jokes were translated, they were presented as if they were originally composed in Russian. This approach intended to steer the respondents’ attention towards evaluating the intensity of the humorous effect, as providing additional information at this stage could have diverted respondents’ focus.

In the second section of the survey, it was revealed to respondents that the jokes were translated. They were offered the original joke in English and asked to identify which translations were generated by AI. One of the multiple-choice options included “All translation variants were produced by a human”.

Automatic evaluations of the translated jokes were performed using online GPT detection tools:

- AI Detector<sup>7</sup>;

- BypassGPT<sup>8</sup> reportedly performing AI detection via GPTzero, Copyleaks, ZeroGPT, Crossplag, Sapling, Writer, Content at Scale;

- GrammarChecker<sup>9</sup>.

<sup>7</sup><https://plagiarismdetector.net/ru/ai-content-detector>

<sup>8</sup><https://bypassgpt.ai/ru>

<sup>9</sup><https://www.grammarchecker.com/ru/ai-text-detector>

## 5. Results and discussion

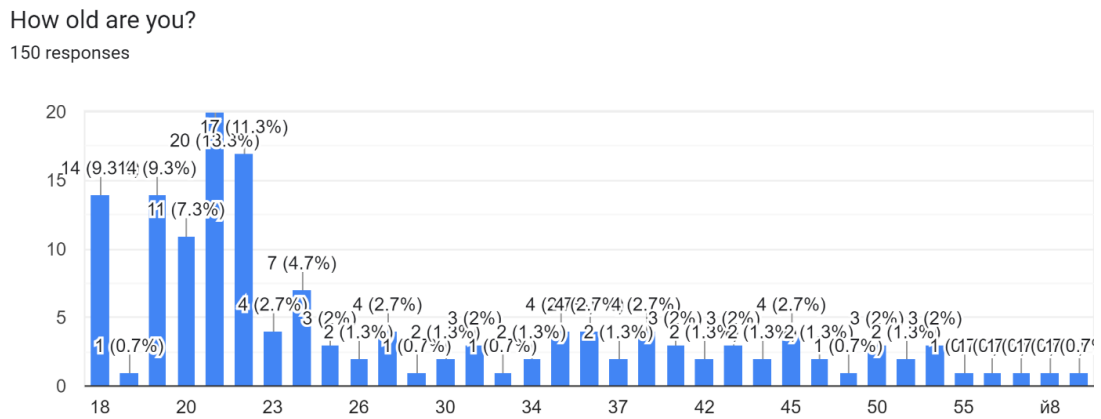
### 5.1. Generalized portrait of human evaluators

The first three figures depict the generalized portrait of the respondents, with the majority being female and within the age range of 18-24. Most respondents evaluated their proficiency in English on a linear scale from 1 to 10, with most rating their proficiency as relatively high (7 to 10). The target group of respondents comprises students and the teaching staff at Don State

Technical University (Rostov-on-Don, Russia). The gender imbalance and age distribution reflect the current demographic situation at this university (68% of female respondents vs 32% of male respondents). Most respondents demonstrated a high self-assessment of their language competence. A linear scale was chosen because not all respondents are familiar with the CEFR framework, which assigns a specific level of language ability.

**Figure 2.** Age distribution among the respondents

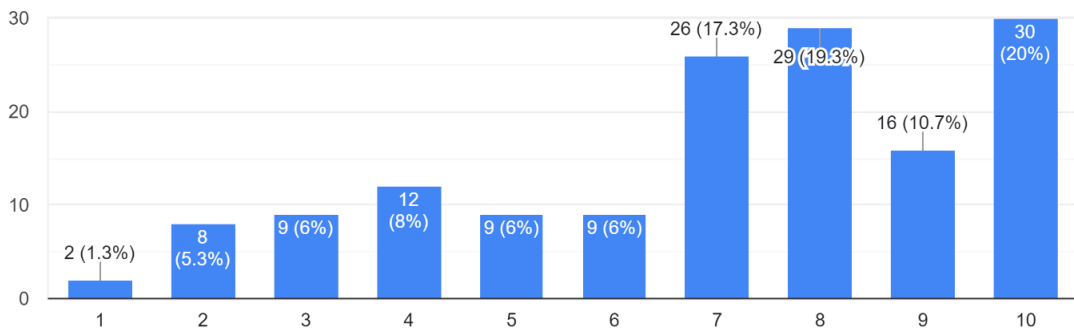
**Рисунок 2.** Распределение респондентов по возрасту



**Figure 3.** Distribution of respondents' perceived command English

**Рисунок 3.** Оценка респондентами собственного уровня владения английским языком

Please rate your command of English on a scale of 1 to 10 (1 being lowest, and 10 being highest).  
150 responses



## 5.2. Joke 1 data

### 5.2.1. Joke 1 translation generation results

Joke 1, in its original English version, is: *'What did the pirate say when he turned 80? Aye matey!'* The wordplay in this joke is based on the homophonic similarity between *'I'm eighty'* and *'Aye matey'*.

The resulting translations for Joke 1, their creators and humour effect mechanism are as follows:

Translation 1 (AI): *'Что сказал пират, когда ему исполнилось семьдесят? Семьдесят футов под килем!'* (semantic exaggeration).

Translation 2 (AI): *'Что сказал пират, когда ему исполнилось сорок? Сорокарабль!'* (lexical blend).

Translation 3 (HT): *'У пирата три тысячи друзей, вот только все они черти.'* (frame-switching, transformation of the clichéd pirate expression, holistic translation technique).

Translation 4 (HT): *'Учительница поставила сыну пирата двойку, и ей пришлось пройтись по доске.'* (holistic translation technique intensified by the play on the double meaning of the word *'доска'*).

When interacting with ChatGPT-4o, we firstly used the prompt template described above. ChatGPT-4o correctly identified the wordplay and suggested a following translation strategy: avoiding a direct translation, considering common Russian pirate phrases, and finding a pun involving age (see Appendix 1<sup>10</sup>). However, the final translation version was unsatisfactory as it led to a complete loss of the pun: *'Что сказал пират, когда ему исполнилось 40? Йо-хо-хо!'* There was a possibility of creating a pun based on the Russian word *'сорок'* by playing on the stereotypical pirate expression that is deeply rooted in Russian popular culture, namely *'Карамба!'*, meaning *'Damn it!'*. The resulting wordplay in this case could be *'Что сказал пират, когда ему исполнилось сорок?'*

*Сорокарамба!'*. The lexical blend *'сорокарамба'* preserves both pirate and age-related themes and has an emotional swashbuckling connotation attached to it, which could be interpreted as *'Damn it, I'm forty and I'm still alive!'* (this phrase evokes a stereotype that a pirate's life is full of danger, and surviving to that age is considered a good luck gift to a pirate). ChatGPT-4o had overlooked the possibility of creating such a pun. We decided against submitting to ChatGPT-4o a human-created translation of the joke as a solution-type to rely on, instead opting for a positive feedback approach by asking ChatGPT-4o to generate two more jokes. The results remained unsatisfactory in terms of preserving the humorous effect: *'Что сказал пират, когда ему исполнилось 50? Йо-пятьдесят!'*, *'Что сказал пират, когда ему исполнилось 60? App-шестьдесят!'*. The first translation version contains a fragment of the pirate phrase *'Yo-ho-ho'*, however, the extent to which it is shortened leaves no room for recognizing it as a specific pirate expression, rendering the translation nonsensical. The second translation version attempts to create a playful sound similarity based on the exclamation *'aargh'*. The New Oxford American dictionary gives the following meaning of *'aargh'*: *"an expression of anguish, horror, rage or other strong emotion, often with humorous intent"*<sup>11</sup>. However, the wide array of emotions expressed by this exclamation is not clearly conveyed in the narrow context of a joke. Furthermore, this phrase is not common in Russian pirate-related culture, and there is no sound similarity between *'App!'* and *'шестьдесят'*, despite ChatGPT-4o's assertion.

Our next step was to provide negative feedback to ChatGPT-4o and ask it to suggest alternative translations, the results were as follows: *'Что сказал пират, когда ему исполнилось 30? Три-дцать!'* and *'Что сказал пират, когда ему исполнилось 40? Сорокарабль!'* The second translation version demonstrates that ChatGPT-4o adopted our

<sup>10</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%201.pdf>

<sup>11</sup> The New Oxford American dictionary (2005). Oxford University Press, New York, USA.

previously mentioned ways of strategizing and created a lexical blend ‘сорокарабль’, which is both pirate- and age-related and somewhat humorous. This translation version was included in the final survey as it aligned with Zabalbeascoa’s binary branching tree structure for translating problem, specifically solution-type 3 (using a different play element and/or imagery) and received 9.3% of respondents’ votes.

However, we sought to elicit more adequate translations from ChatGPT-4o and asked it to clarify the previous response, particularly the joke containing the Russian exclamation ‘App!’ as there was no transparent meaning attached to it. ChatGPT-4o provided a general explanation of the meaning without specifying any particular meaning used in translation. Additionally, it suggested two other translation versions of the joke: ‘Что сказал пират, когда ему исполнилось 70? Семьдесят футов под килем!’ and ‘Что сказал пират, когда ему исполнилось 60?/

Йо-шестьдесят-хо!’. This time, ChatGPT-4o took a common nautical blessing ‘семь футов под килем’ and transformed the numeral 7 into 70, creating a humorous effect by exaggerating the blessing’s power. According to Zabalbeascoa’s binary branching tree structure, this solution falls into type 3, which involves using a different play element and/or imagery. This translation version was also included in the final survey and was evaluated as humorous by 13.3% of respondents (see Figure 4).

Two other translation versions are human-created: ‘У пирата три тысячи друзей, вот только все они черти’ and ‘Учительница поставила сыну пирата двойку, и ей пришлось пройтись по доске’. They completely rearranged the scenario yet maintained metadata of the joke: pirate expressions, realia and numerals. This approach, known as the holistic translation technique, produced the highest humorous effect (see Figure 4). These jokes received 51.3% and 26% of respondents’ votes.

**Figure 4.** Joke 1 human evaluation ratings for humorous effect

**Рисунок 4.** Оценка респондентами степени выраженности юмористического эффекта в шутке 1

Which joke do you think is the funniest?  
150 responses



Linguistic features in Joke 1 contain text parameters which make it possible to note that AI-generated translations are more similar to the original English joke in all the enlisted parameters while human-generated translation contain vivid deviations from the source text. It can be assumed that sentence

distribution and length as well as punctuation can be important evidence of the holistic approach used by a human translator (see Table 1<sup>12</sup>).

<sup>12</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%201.%20Linguistic%20features%20of%20Joke%201%20and%20its%20Russian%20translations.pdf>

Repetitiveness analysis shows a similar number of 1-grams in all jokes, with human-translated jokes still being closer to the source text. Nevertheless, word frequency assessment performed has demonstrated that human translators tend to use more difficult words and notions.

Readability scores show that there is only one translation which can be considered an equivalent for this category – Translation 3 performed by a human translator (see Figure 7). Both versions of Joke 1 are good for Grade 1 6/7-year-old children. As far as other translations are concerned, they are

designed for a more senior age group, including 14-year-old teenagers.

According to POS scores, human translations again are statistically closer to the English joke, with Translation 3 being the most equivalent.

Finally, as can be seen from Table 1, neither AI nor human translators were able to use the phonemic-level humour effect in Russian translations. Another challenge is the humour translation solution, which points at the impossibility of humorous element and/or imagery preservation.

**Table 2.** Automatic evaluation results for Joke 1

**Таблица 2.** Результаты автоматической оценки для шутки 1

AI Detector Tool	Translation 1 (AI)	Translation 2 (AI)	Translation 3 (HT)	Translation 4 (HT)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	100% AI 0% HT	100% AI 0% HT	48% AI 52% HT	100% AI 0% HT

### 5.2.2. Joke 1 evaluation results

In the second part of our survey, respondents were provided with the original joke in English and asked to identify which of the translations were generated by AI. Among the multiple-choice options there was also an option “All translation variants were produced by a human”.

The results presented in Figure 5 show that the majority of respondents successfully differentiate between human and AI-generated content: the first and second translation variants were correctly identified as AI translations by 35.3% and 41.3% of

respondents, respectively. Additionally, respondents were asked to identify specific words or expressions in the translations of Joke 1 that suggested the text might have been generated by AI, the most frequently cited indicators were the following: ‘сорокарабль’, ‘семьдесят футов под килем’, ‘что сказал пират’, literal translation.

The results also demonstrate that many respondents had difficulty distinguishing between human and AI-generated translations: 35.3% were unable to make this distinction.

**Figure 5.** Respondents' identification of AI-generated translations in Joke 1

**Рисунок 5.** Обнаружение респондентами признаков ИИ в переводах шутки 1

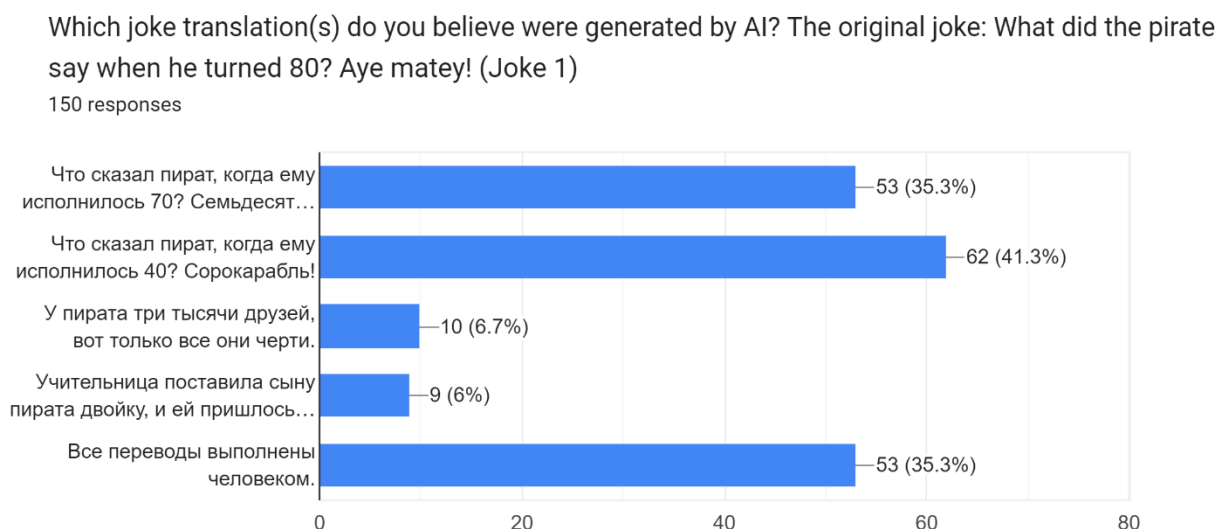


Table 2 contains the results of automatic evaluation for Joke 1 performed via three different online tools. As can be seen in this and other similar tables, the first two tools turn out to be inconsistent in their evaluations because they label all texts as AI-generated. Only the third tool shows more adequate results, which nevertheless are not reliable either as they are not stable enough.

### 5.3. Joke 2 data

#### 5.3.1. Joke 2 translation generation results

The original version of Joke 2 is: 'Candy is dandy but liquor is quicker'. This short poem, composed by the American humorist Ogden Nash in his work *Reflections on Ice-breaking* (1931), was recited by Willy Wonka in the 1971 musical fantasy film *Willy Wonka & the Chocolate Factory*. However, in the Russian audiovisual translation, the original wordplay was lost, resulting in dialogue that appeared absurd:

*Mr. Salt: [noticing signs on vats] Wonka. Butterscotch? Buttergin? Got a little something going on the side?*

*Willy Wonka: Candy is dandy, but liquor is quicker*<sup>13</sup>.

Russian translation:

– Ириски? Сливки? Внутри что-то происходит?

– Превосходный леденец.

The resulting translations for Joke 2, their creators and humour effect mechanism are as follows:

Translation 1 (HT): 'Конфеты – эффектны, но аперитивы – оперативней!' (conflicting schemas, two opposed scripts (love/courtship vs food) combined in the joke, reinforced by rhyming).

Translation 2 (AI): 'Пирожок – хорош, но коньяк – ништяк!' (conflicting schemas, script opposition (good food vs bad food) reinforced by rhyming).

Translation 3 (AI): 'Конфеты – это мило, но спиртное – это сила.' (conflicting schemas, script opposition (good food vs bad food) reinforced by rhyming).

Translation 4 (AI): 'Конфета – для привета, а водка – для разгона.'

<sup>13</sup> Willy Wonka & the Chocolate Factory. Directed by Mel Stuart, Wolper Pictures Ltd., 1971. Warner Bros. Entertainment.

(conflicting schemas, script opposition (good food vs bad food) reinforced by rhyming).

Using the cross-lingual thought prompting method, we asked ChatGPT-4o to translate Joke 2 (see Appendix 2<sup>14</sup>). However, this translation version exhibited a misinterpretation: ChatGPT-4o incorrectly identified the rhyme between 'dandy' and 'quicker' and the comparison between something sweet and something strong or fast, as the key elements constituting the essence of the joke. Notably, there is no rhyme between 'dandy' and 'quicker', moreover, the joke implies the addressee's knowledge of the extralinguistic world, as tertium comparationis linking 'dandy' and 'quicker' is founded on social and cultural presuppositions. Overlooking this interplay of purely linguistic and empirical aspects resulted in a poor translation that lacked both humorous effect and rhyme: 'Конфеты – это мило, а спиртное – мобильно'. Consequently, we employed an explanatory technique combined with positive feedback approach: *Not bad, but I need a more pronounced humorous effect. The meaning of the joke is that candy is good for elegant courtship, but if you are looking for a speedy seduction, liquor is better. Try again.* The resulting translation was satisfactory, though the language register shifted toward a neutral tone ('Конфеты – это мило, но спиртное – это сила.'). As illustrated in Figure 6, this translation ranked second, following the human-created translation ('Конфеты – эффективны, но aperитивы – оперативней!'), which was evaluated as humorous by 34% of respondents. It should

also be noted that while translating Joke 2, the students discovered a pre-existing translation by A. Zhukov for the English-Russian Encyclopedic Dictionary *AMERICANA*<sup>15</sup>. This translation discouraged them from submitting their own work, which they self-evaluated as 'unsatisfactory', so this is the only translation variant not originally produced by the students.

Two additional translation versions, generated by ChatGPT-4o ('Пирожок – хорош, но коньяк – ништяк', 'Конфета – для привета, а водка – для разгона') were elicited from a new dialogue. In this case, an explanatory technique was not employed, which led to a shift in the language register, the final translation acquired slang-to-vulgar tone (ништяк) and utilized stereotypical cultural references associated with Russian culture ('пирожок', 'водка'). According to Zabalbeascoa's binary branching tree structure, all AI-generated translations fall into solution type 3, which involves partial loss of the initial play elements.

Table 3<sup>16</sup> presents linguistic features for Joke 2. Interestingly, the human-translated Russian version is textometrically different from the source text in terms of emotional semantics and readability. Nevertheless, the terminological and structural complexity of this translation does not reduce the humorous effect and accessibility of the text. At the same time, such complexity might be considered a necessary compromise to preserve the initial play element and imagery, which makes this version of joke 2 the most popular among the respondents.

<sup>14</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%202.pdf>

<sup>15</sup> Chernov, Ghelly V. (Ed). (1996) *AMERICANA. English-Russian Encyclopedic Dictionary*, Polygramma, Smolensk, Russia.

<sup>16</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%203.%20Linguistic%20features%20of%20Joke%20and%20its%20Russian%20translation.pdf>



**Figure 6.** Joke 2 human evaluation ratings for humorous effect

**Рисунок 6.** Оценка респондентами степени выраженности юмористического эффекта в шутке 2

Which joke do you think is the funniest?

150 responses



### 5.3.2. Joke 2 evaluation results

The results presented in Figure 7 demonstrate that the only human-translated joke was incorrectly identified as AI-generated by 42% of respondents, while AI-generated translations 2–5 exhibited lower identification rates in comparison to the human-translated version 1. The following

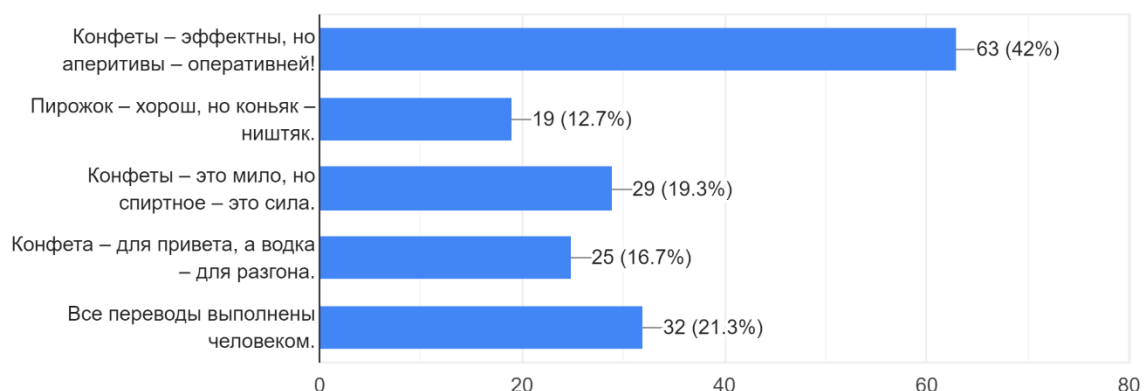
words and expressions were highlighted as AI-like elements: *эффектны, оперативней, разгон, пирожок, пирожок – хорош, ништяк, конфеты – это мило спиртное, particle 'это', literal translation.* Human evaluation can be also compared to automatic AI evaluation (see Table 4).

**Figure 7.** Respondents' identification of AI-generated translations in Joke 2

**Рисунок 7.** Обнаружение респондентами признаков ИИ в переводах шутки 2

Which joke translation(s) do you believe were generated by AI? The original joke: Candy is dandy but liquor is quicker. (Joke 2)

150 responses



**Table 4.** Automatic evaluation results for Joke 2

**Таблица 4.** Результаты автоматической оценки для шутки 2

AI Detector Tool	Translation 1 (HT)	Translation 2 (AI)	Translation 3 (AI)	Translation 4 (AI)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	99% AI 1% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT

#### 5.4. Joke 3 data

##### 5.4.1. Joke 3 translation generation results

The original version of Joke 3 is: *'I took the shell off my racing snail, thinking it would make him run faster. If anything, it made him more sluggish.'* The linguistic mechanism underlying this joke is based on assigning a new contextual meaning to the word *'sluggish'*, interpreting it as *'like a slug'*. According to Collins Dictionary, the suffix *-ish* *'is added to nouns and names to form adjectives which indicate that someone or something is like a particular kind of person or thing.'*<sup>17</sup> The logical reasoning behind the joke is as follows:

1. A snail's shell is perceived as a burden that hinders its ability to move faster.

2. Removing a shell transforms a snail into an ordinary slug, contrary to the expectation that this procedure will speed it up.

The word *'sluggish'* also has a fixed meaning: *'You can describe something as sluggish if it moves, works, or reacts much slower than you would like or is normal.'* The humorous effect is achieved through the simultaneous realization of two meanings – the fixed, literal meaning and the new, contextual meaning – reinforced by the concept of snail racing.

The resulting translations for Joke 3, their creators and humour effect mechanism are as follows:

Translation 1 (AI): *'Я снял раковину со своей гоночной улитки, думая, что это ускорит ее. На самом деле, она стала еще более слизнявой.'* (semantic ambiguity).

Translation 2 (AI): *'Я снял раковину с моей гоночной улитки, думая, что она станет быстрее. Но в итоге она стала не спринтером, а сползнером.'* (lexical blend).

Translation 3 (HT): *'Улитка Ульяна ползла на работу так быстро, что выползла из домика.'* (script opposition reinforced by vocalic alliteration).

Translation 4 (HT): *'Я решил поменять панцирь на своей гоночной улитке, надеясь ускорить её. В итоге она застряла на пит-стопе.'* (script opposition).

Translation 5 (HT): *'Надеясь сделать улитку быстрее, я снял с неё раковину. Но на улиточных гонках её дисквалифицировали.'* (script opposition).

To maintain the humour, ChatGPT-4o incorporated the word *'слизнявый'* into the translation, leaving the original syntactic structure intact: *'Я снял раковину со своей гоночной улитки, думая, что это ускорит её. На самом деле, она стала еще более слизнявой.'* (see Appendix 3<sup>18</sup>). However, it should be pointed out that the adjective *'слизнявый'* carries a disdainful connotation

<sup>17</sup> Collins English Dictionary (2024), available at: <https://www.collinsdictionary.com/dictionary/english/sluggish> (Accessed 20 July 2024).

<sup>18</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%203.pdf>

in Russian, denoting a miserable or worthless person (слизняк). A stylistically neutral adjective derived from 'слизняк' is 'слизневый'. In the context of the snail's characteristics, 'слизнявый' has the potential to create a humorous effect by anthropomorphizing the snail; additionally, 'слизнявый' can be interpreted as 'pertaining to or characteristic of a slug' within the context of the joke. Thus, this translation version employs the same linguistic mechanism of playing on the dual meaning as the original joke, although the humorous effect is less pronounced, resonating with only 8.7% of respondents, as shown in Figure 8.

The second translation by ChatGPT-4o utilized the technique of creating a new lexical blend from the words 'сползать' and 'спринтер', enhancing a humorous effect through semantic contradiction arising from

such a blend ('сползать' conveys the idea of slow movement, while 'спринтер' denotes fast movement): 'Я снял раковину с моей гоночной улитки, думая, что она станет быстрее. Но в итоге она стала не спринтером, а сползнером.' This joke was evaluated as humorous by 13.3% respondents (see Figure 8).

The joke's human translation 'Улитка Ульяна ползла на работу так быстро, что выползла из домика' was evaluated as the most humorous by respondents (42%), two other human translations 'Я решил поменять панцирь на своей гоночной улитке, надеюсь ускорить её. В итоге она застряла на пистоне', and 'Надеясь сделать улитку быстрее, я снял с неё раковину. Но на улиточных гонках её дисквалифицировали' received 14.7% and 21.3% of respondents' votes, respectively (see Figure 8).

**Figure 8.** Joke 3 human evaluation ratings for humorous effect

**Рисунок 8.** Оценка респондентами степени выраженности юмористического эффекта в шутке 3

Which joke do you think is the funniest?

150 responses



Linguistic feature of Joke 3 demonstrate that AI-generated Russian translations are structurally and stereotypically closer to the source text, while human-translated Russian jokes are characterized by a sound holistic approach (see Table 5<sup>19</sup>). Although both AI-

and human-generated translations manage to maintain solution type 2 (partial loss of the initial play elements), only human translations have scored maximum results.

<sup>19</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%205.%20Linguist>

ic% 20features% 20of% 20Joke% 203% 20and% 20its% 20Russian% 20translations.pdf

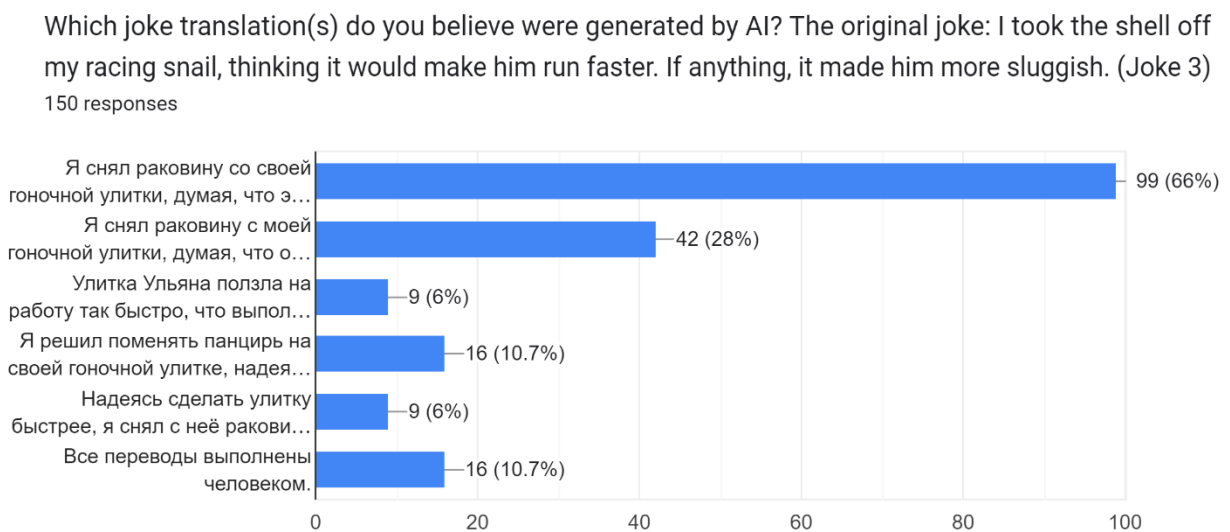
### 5.4.2. Joke 3 evaluation results

The results presented in Figure 9 reveal that the majority of respondents correctly identified the first and second translation variants as AI-generated, however, the identification rate for the second AI-translation significantly decreased to 28% compared to the first AI translation, which had 66% identification rate. The following

words and expressions were noted as disrupting the perception of human-like translation: ‘слизнявая’, ‘сползнер’, ‘я снял раковину’, ‘гоночная улитка’, ‘панцирь’, ‘nut-stop’, literal translation, detailed translation. Human evaluation can be also compared to automatic AI evaluation (see Table 6).

**Figure 9.** Respondents’ identification of AI-generated translations in Joke 3

**Рисунок 9.** Обнаружение респондентами признаков ИИ в переводах шутки 3



**Table 6.** Automatic evaluation results for Joke 3

**Таблица 6.** Результаты автоматической оценки для шутки 3

AI Detector Tool	Translation 1 (AI)	Translation 2 (AI)	Translation 3 (HT)	Translation 4 (HT)	Translation 5 (HT)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	100% AI 0% HT	100% AI 0% HT	99% AI 1% HT	0% AI 100% HT	99% AI 1% HT

### 5.5. Joke 4 data

#### 5.5.1. Joke 4 translation generation results

The original version of Joke 4 is: ‘Two windmills are standing in a field and one asks,

“What’s your favorite kind of music?” The other says, “I’m a big metal fan”.’ The wordplay in this joke is activated by the double meaning of the collocation ‘metal fan’, which can be interpreted as ‘a fan of metal music’ and ‘flat object that you hold in your

*hand and wave in order to move the air and make yourself feel cooler.*<sup>20</sup>

The resulting translations for Joke 4, their creators and humour effect mechanism are as follows:

Translation 1 (AI): *‘Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты любишь?» Второй отвечает: «Я обожаю бриз-энд-ролл!»* (lexical blend).

Translation 2 (HT): *‘Одна мельница говорит другой: «Как ты могла не поздравить меня с днём рождения?!» А та ей отвечает: «Извини, что-то совсем закрутилась»* (semantic ambiguity).

Translation 3 (AI): *‘Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты любишь?» Второй отвечает: «Мне нравится ветер-рок!»* (lexical blend).

Translation 4 (HT): *‘Стоят две мельницы на ферме. Одна у другой спрашивает: «Какой у тебя любимый жанр музыки?» А та ей отвечает: «Кантри».* (semantic ambiguity).

ChatGPT-4o correctly analyzed the joke’s mechanism, and aimed to find an equivalent wordplay in Russian that fits the context and maintains a humour (see Appendix 4<sup>21</sup>). The result was as follows: *‘Два ветряка стоят в поле, и один спрашивает: «Какая твоя любимая музыка?» Второй отвечает: «Я большой поклонник тяжелого металла».* The joke is based on the extralinguistic knowledge that wind turbines are typically made from metals and weigh many tons. ChatGPT-4o employed metonymic transfer founded on the humorous cause-and-effect relationship to create a humorous effect: the wind turbine is made from metal and weighs many tons, and therefore it might appreciate heavy metal music. Although this translation can be

considered humorous, the joke seems somewhat contrived and does not receive an instant humorous response. Consequently, we asked ChatGPT-4o to refine the answer by using encouraging and explanatory techniques, the results obtained were included in the survey: *‘Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты любишь?» Второй отвечает: «Мне нравится ветер-рок!»* and *‘Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты любишь?» Второй отвечает: «Я обожаю бриз-энд-ролл!»* In both jokes ChatGPT-4o preserved the original pun by incorporating novel hyphenated compound words ‘ветер-рок’ and ‘бриз-энд-ролл’ into the joke. The novel compound ‘ветер-рок’ is homophonic with ‘ветерок’, meaning ‘little wind’ in Russian, which activates double meaning, and the rolling ‘r’ at the stem-junction emphasizes the playfulness. The novel compound ‘бриз-энд-ролл’ also demonstrates ChatGPT-4o’s contextually aware approach by retaining the ‘wind theme’ while simultaneously keeping the music genre identifiable. According to Zabalbeascoa’s binary branching tree structure, these AI-generated translations fall into solution type 2, which involves partial loss of the initial play elements. The ChatGPT-4o’s translation received 12% and 10% of respondents’ votes, respectively (see Figure 10).

The human translation, which produced the highest humorous effect, employed a holistic translation technique (solution type 3), retaining the participants of a communicative situation, but changing the situation itself by exploiting double meaning of the word ‘закрутиться’: *‘Одна мельница говорит другой: «Как ты могла не поздравить меня с днём рождения?!» А та ей отвечает: «Извини, что-то совсем закрутилась»*. This joke was evaluated as humorous by 62% of respondents.

Another human-created translation of the joke, which closely adheres to the original joke’s form and content, slightly modified the setting by incorporating a ‘farm’ and changing

<sup>20</sup> Collins English Dictionary (2024), available at: <https://www.collinsdictionary.com/dictionary/english/s-luggish> (Accessed 20 July 2024).

<sup>21</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%204.pdf>

the music genre to 'country', logically inferred from the new setting: 'Стоят две мельницы на ферме. Одна у другой спрашивает: «Какой у тебя любимый жанр музыки?» А та ей отвечает:

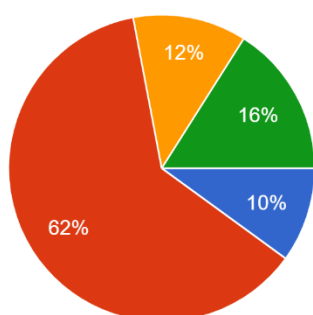
«Кантри».' This human-created joke translation falls into solution type 2 and was evaluated as humorous by 16% of respondents (see Figure 10).

**Figure 10.** Joke 4 human evaluation ratings for humorous effect

**Рисунок 10.** Оценка респондентами степени выраженности юмористического эффекта в шутке 4

Which joke do you think is the funniest?

150 responses



- Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты любишь?» Второй отвечает: «Я обо...
- Одна мельница говорит другой: «Как ты могла не поздравить меня с днём рождения?!» А та ей отвечает: «Изв...
- Два ветряка стоят в поле. Один спрашивает: «Какую музыку ты люб...
- Стоят две мельницы на ферме. Одна у другой спрашивает: «Какой у тебя...

Joke 4 features properties which demonstrate contrastive differences between AI-generated and human-translated versions of the joke (see Table 7<sup>22</sup>). Calques and literal translation solutions create a psychological obstacle for the interpretation of jokes. Besides, the statistics shows that AI-generated texts tend to have higher and more positive sentiment score than human translations while autosemanticity tends to be lower. This phenomenon is supported by our empirical results obtained via human evaluation where the respondents give corresponding comments.

### 5.5.2. Joke 4 evaluation results

The results shown in Figure 11 indicate that 30.7% and 34% of respondents correctly identified jokes generated by AI; however, the majority (35.3%) believed that all jokes were translated by a human. Interestingly, the joke rated as the most humorous received the lowest rate of AI attribution, with only 5.3% of respondents attributing it to AI. The most recurrent words and patterns, identified as indicators of AI translation, included: 'бриз-энд-ролл', 'ветер-рок', 'ветряк', 'кантри' and a syntactic structure 'два ветряка стоят в поле'. Human evaluation can be also compared to automatic AI evaluation (see Table 8).

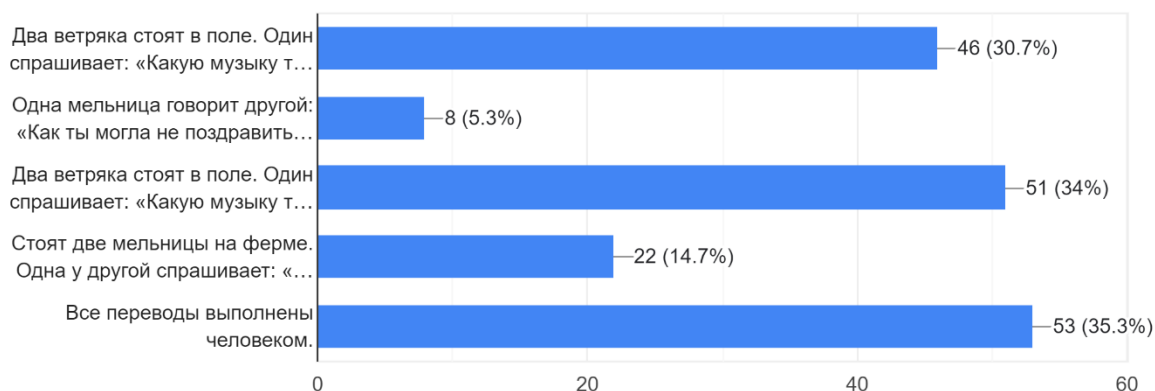
<sup>22</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%207.%20Linguistic%20features%20of%20Joke%204%20and%20its%20Russian%20translations.pdf>

**Figure 11.** Respondents' identification of AI-generated translations in Joke 4

**Рисунок 11.** Обнаружение респондентами признаков ИИ в переводах шутки 4

Which joke translation(s) do you believe were generated by AI? The original joke: Two windmills are standing in a field and one asks, "What's your favori...sic?" The other says, "I'm a big metal fan." (Joke 4)

150 responses



**Table 8.** Automatic evaluation results for Joke 4

**Таблица 8.** Результаты автоматической оценки для шутки 4

AI Detector Tool	Translation 1 (AI)	Translation 2 (HT)	Translation 3 (AI)	Translation 4 (HT)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	37% AI 63% HT	99% AI 1% HT	0% AI 100% HT	99% AI 1% HT

## 5.6. Joke 5 data

### 5.6.1. Joke 5 translation generation results

The original version of Joke 5 is: “‘I have a split personality,” said Tom, being frank.’ The linguistic mechanism underlying the joke relies on the double meaning of the word ‘frank’, which can refer to the adjective meaning ‘sincere’ or function as a proper name.

The resulting translations for Joke 5, their creators and humour effect mechanism are as follows:

Translation 1 (AI): ‘У меня раздвоение личности, – сказал Том, будучи Франком.’ (semantic ambiguity).

Translation 2 (HT): ‘Кажется, у меня раздвоение личности, – сказал Том, сидя перед зеркалом.’ (script opposition).

Translation 3 (AI): “‘У меня раздвоение личности, – сказал Том, и его второе «я» кивнуло в поддержку.’ (script opposition).

Translation 4 (HT): ‘– Кажется, у меня раздвоение личности, – сказал Том, глядясь в зеркало. – У нас, – мягко поправило отражение.’ (script opposition).

In its first translation, ChatGPT-4o did not make any alterations to the joke resorting to a literal translation (see Appendix 5<sup>23</sup>). The

<sup>23</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%205.pdf>

df

final translation was as follows: ‘«У меня раздвоение личности», – сказал Том, будучи Франком.’ Strange as it may seem, this translation retains some potential for achieving humorous effect, as the proper name ‘Франк’ may be interpreted in Russian as a currency (Swiss Franc); the humorous effect is further intensified by a converb ‘будучи’, a literary form that is perceived as obsolete. However, the humorous effect of this translation version is not highly satisfactory, which is why we asked ChatGPT-4o to change the names and scenario while adhering to the theme of a split personality. On the third attempt, we obtained a satisfactory result: ‘«У меня раздвоение личности», – сказал Том, и его второе “я” кивнуло в поддержку.’. These AI-generated jokes were evaluated as humorous by 12.7% and 18.7% of respondents, respectively (see Figure 12).

The human-created joke, which received 10.7% of respondents’ votes, is quite similar in its translation approach to the previously discussed one: «Кажется, у меня раздвоение личности», – сказал Том, сидя перед зеркалом.’ It omits the name Frank and introduces the utterance in a new context. However, this is the only instance where a human-created joke ranked third after an AI-generated joke in terms of achieving a humorous effect (see Figure 12).

The highest humorous effect (58%) was achieved by refining the human-created joke and enhancing it in terms of scenario by creating a new punchline: ‘– Кажется, у меня раздвоение личности, – сказал Том, глядясь в зеркало. – У нас, – мягко поправило отражение’ (see Figure 12).

All translations (both human-created and AI-generated) fall into solution type 2 (partial loss of the initial play elements).

**Figure 12.** Joke 5 human evaluation ratings for humorous effect

**Рисунок 12.** Оценка респондентами степени выраженности юмористического эффекта в шутке 5

Which joke do you think is the funniest?  
 150 responses



As can be seen in Table 9<sup>24</sup>, linguistics features for English and Russian versions of Joke 5 show that AI-generated translations deviate from the source text textometric parameters more than human translations and

feature more structural complexity. Human translations, though also deviating in some categories, demonstrate more translation holisticity and repetitiveness efficiency.

### 5.6.2. Joke 5 evaluation results

Figure 13 reveals that 67.3% of respondents correctly identified the first translation variant as AI-generated. In contrast, the third AI-generated translation

<sup>24</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%209.%20Linguistic%20features%20of%20Joke%205%20and%20its%20Russian%20translations.pdf>

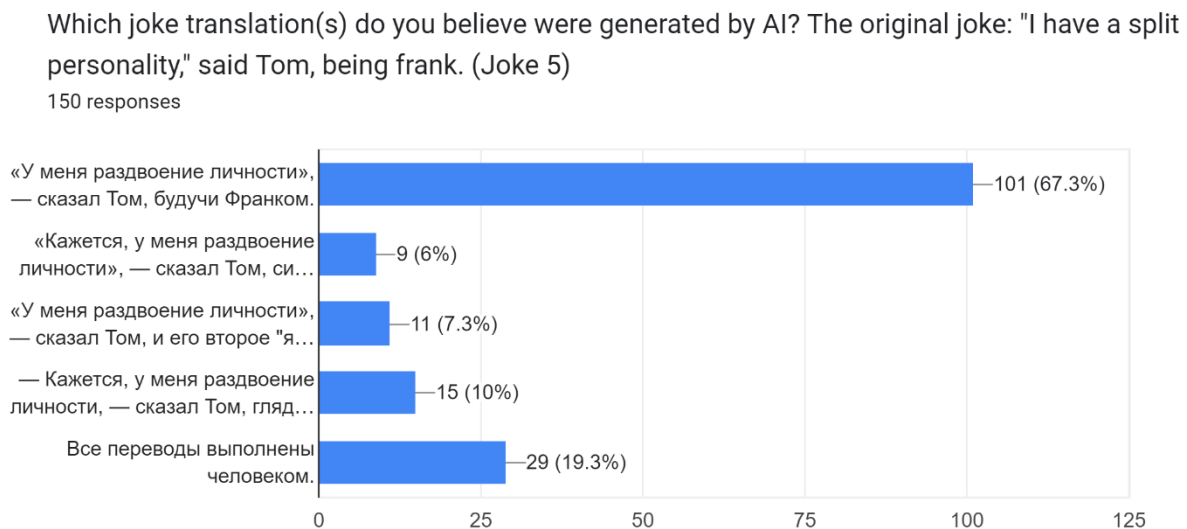


had a lower rate of correct classification. A moderate number of respondents (19.3%), considered all options to be human-created. The most frequently mentioned words and expressions that respondents believed indicated AI presence in the translation were: ‘будучи Франком’, ‘кивнуло в поддержку’,

‘второе “я”’. Respondents also pointed out uncomfortable extreme literalness, which they associated with translations produced by AI without singling out specific grammatical structures and lexis. Human evaluation can be also compared to automatic AI evaluation (see Table 10).

**Figure 13.** Respondents’ identification of AI-generated translations in Joke 5

**Рисунок 13.** Обнаружение респондентами признаков ИИ в переводах шутки 5



**Table 10.** Automatic evaluation results for Joke 5

**Таблица 10.** Результаты автоматической оценки для шутки 5

AI Detector Tool	Translation 1 (AI)	Translation 2 (HT)	Translation 3 (AI)	Translation 4 (HT)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	98% AI 2% HT	99% AI 1% HT	99% AI 1% HT	99% AI 1% HT

## 5.7. Joke 6 data

### 5.7.1. Joke 6 translation generation results

The original version of Joke 6 is: ‘I tried catching fog yesterday. Mist.’ The comic base of this joke is activated by the linguistic incongruity between spelling and sound: ‘mist’, referring to the physical condensation, clashes with ‘missed’, meaning failing to achieve something.

The resulting translations for Joke 6, their creators and humour effect mechanism are as follows:

Translation 1 (AI): ‘Что гроза сказала молнии? «Ты меня всегда поражаешь!»’ (semantic ambiguity).

Translation 2 (AI): ‘Почему ветер не приходит на свидания? Потому что он всегда «сдувается» в последний момент!’ (semantic ambiguity).

Translation 3 (AI): *‘Почему дождь не рассказывает анекдоты? Потому что у него всегда плохое настроение!’* (script opposition).

Translation 4 (HT): *‘Я вчера гулял по Лондону, но мало что помню... Всё, как в тумане.’* (semantic ambiguity).

Translation 5 (AI): *‘Вчера пытался поймать туман. Промазал.’* (semantic ambiguity).

ChatGPT-4o correctly analyzed the joke’s mechanism and employed a strategy of using a different type of wordplay in Russian that involves similar meanings or sounds (see Appendix 6<sup>25</sup>). The result was as follows: *‘Вчера пытался поймать туман. Промазал.’* ChatGPT-4o’s translation is a syntactic calque from Russian, where form has been severed from content. ChatGPT-4o argues that this translation maintains the idea of trying to catch fog and plays on *‘промазал’* which can be understood as failing to catch something. We hold the view that the translation of the joke acquired an air of absurdity, because the situation itself is incongruent with our knowledge of the external world. Given the unsatisfactory result, we asked ChatGPT-4o to replace *‘туман’* with a new word that creates a pun on *‘промазал’*, the resulting joke was: *‘Вчера пытался намазать масло. Промазал.’* This version was included in the final survey and received 30% of respondents’ votes (see Figure 14).

To elicit additional translations from ChatGPT-4o, we employed a holistic translation technique and requested that ChatGPT-4o create a series of jokes based on different weather conditions to produce a humorous effect in Russian. The following results were satisfactory: (1) *‘Что гроза сказала молнии? “Ты меня всегда поражаешь!”’*, (2) *‘Почему ветер не приходит на свидания? Потому что он всегда ‘сдувается’ в последний момент!’*

and (3) *‘Почему дождь не рассказывает анекдоты? Потому что у него всегда плохое настроение!’* (see Appendix 6) ChatGPT-4o performed the task relatively successfully, generating pun-like question-and-answer jokes that animate natural phenomena. The joke’s translation versions (1) and (2) are based on semantic ambiguity activated by words *‘поражаешь’* and *‘сдуваться’*, while the translation version (3) is de-punned and plays on the negative cultural connotations of the word *‘rain’* which is often associated with tears and sadness. These jokes received low funniness ratings: (1) – 16%, (2) – 15.3%, (3) – 4.7%, respectively (see Figure 14).

The human-translated joke received the highest ranking (34%): *‘Я вчера гулял по Лондону, но мало что помню... Всё как в тумане.’* The joke accomplishes the humorous effect by incorporating a new punchline *‘Всё как в тумане’*, in which two different underlying semantic structures (literal and figurative) are presented by a single surface structure.

All translations (both human-created and AI-generated) fall into solution type 3 (using a different play element/imagery).

Linguistic features of Joke 6 in its English and Russian versions maintain the same trend to meet lower complexity and higher readability expectations of respondents (see Table 11<sup>26</sup>). Apart from the obvious translation challenge concerning word frequency ratings and terminological complexity in processing Russian translations of Joke 6, these parameters were not critical for the respondents.

<sup>25</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%206.pdf>

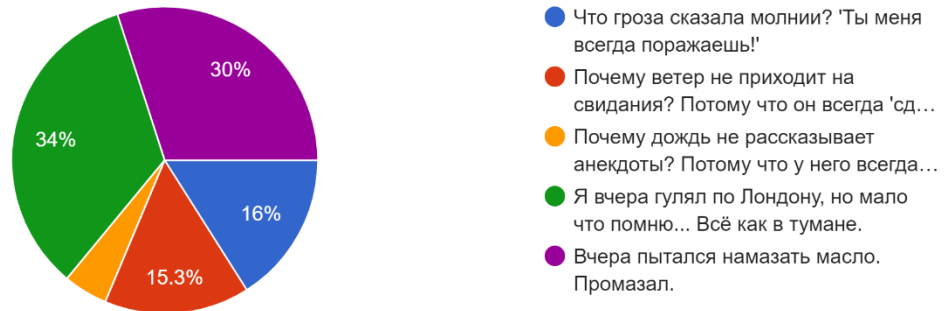
<sup>26</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%2011.%20Linguistic%20features%20of%20Joke%206%20and%20its%20Russian%20translations.pdf>

**Figure 14.** Joke 6 human evaluation ratings for humorous effect

**Рисунок 14.** Оценка респондентами степени выраженности юмористического эффекта в шутке 6

Which joke do you think is the funniest?

150 responses



### 5.7.2. Joke 6 evaluation results

The results shown in Figure 15 indicate that the highest percentage of respondents (44%) perceived all translation variants as being generated by a human, whereas the only human-created translation was mistakenly perceived as AI-generated by 18% of

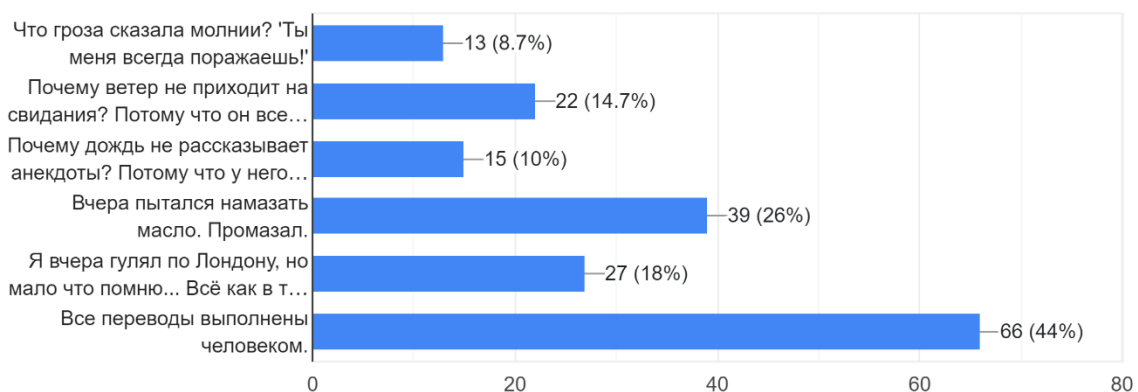
respondents. The words and expressions most frequently identified by respondents as indicative of AI presence in the translation were: 'промазал', 'сдувается', 'всё как в тумане', 'мало что помню'. Human evaluation can be also compared to automatic AI evaluation (see Table 12).

**Figure 15.** Respondents' identification of AI-generated translations in Joke 6

**Рисунок 15.** Обнаружение респондентами признаков ИИ в переводах шутки 6

Which joke translation(s) do you believe were generated by AI? The original joke: I tried catching fog yesterday. Mist. (Joke 6)

150 responses



**Table 12.** Automatic evaluation results for Joke 6

**Таблица 12.** Результаты автоматической оценки для шутки 6

AI Detector Tool	Translation 1 (AI)	Translation 2 (AI)	Translation 3 (AI)	Translation 4 (HT)	Translation 5 (AI)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
Grammar Checker	100% AI 0% HT	0% AI 100% HT	99% AI 1% HT	92% AI 8% HT	46% AI 54% HT

## 5.8. Joke 7 data

### 5.8.1. Joke 7 translation generation results

The original version of Joke 7 is: *'I can't believe I got fired from the calendar factory. All I did was take a day off.'* The humour in this joke is based on using a phrase with double meaning in a specific context where both meaning potentialities are realized. Within the context of a calendar factory, the phrase 'take a day off' could be interpreted literally as 'removing a day from the calendar', which would be perceived as unsatisfactory job performance; it also refers to taking a one day's break from work, which is the normal practice. The comic effect is evoked by the juxtaposition of these two meanings.

The resulting translations for Joke 7, their creators and humour effect mechanism are as follows:

Translation 1 (HT): *'Газманова уволили с фабрики по производству календарей, потому что он оставлял ясные дни себе.'* (script opposition).

Translation 2 (AI): *'Невероятно, но факт: уволили меня с завода по производству календарей. Причина? Один пропущенный день.'* (semantic ambiguity).

Translation 3 (AI): *'Не могу поверить, что меня уволили с фабрики календарей. Я всего лишь взял один день отгула.'* (semantic ambiguity).

ChatGPT-4o's translation strategy involved avoiding the phrase *'взять выходной'*, the direct translation of *'take a*

*day off'*, as it does not inherently contain the double meaning that plays with the idea of removing a day from a calendar (see Appendix 7<sup>27</sup>). In an attempt to find a Russian equivalent that conveys both meanings, ChatGPT-4o played with the word *'день'*, generating the following joke: *'Не могу поверить, что меня уволили с фабрики календарей. Я всего лишь взял один день отгула.'* Although ChatGPT-4o suggests that a pun is created on the word *'day'*, it is the use of a verb *'взять'* that creates a pun, as it can mean both physically taking something away (tearing a day out of a calendar) and taking a day off in a figurative sense. However, this translation does not produce a pronounced humorous effect, as the verb *'взять'* functions as part of a support verb construction and is thus desemantized (empty of meaning). Consequently, we requested ChatGPT-4o to approach the translation creatively in order to intensify the humorous effect, the result was as follows: *'Невероятно, но факт: уволили меня с завода по производству календарей. Причина? Один пропущенный день.'* In this translation, ChatGPT-4o structured the joke around the phrase *'пропущенный день'*, which can be interpreted both as *'accidentally skipping a day while compiling a calendar'* and *'missing a day of work'*. Both AI-

<sup>27</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%207.pdf>

generated translations fall into solution type 1 (maintaining the initial play element and/or imagery). These translation versions were included in the final survey and received 10% and 34% of respondents' votes, respectively (see Figure 16).

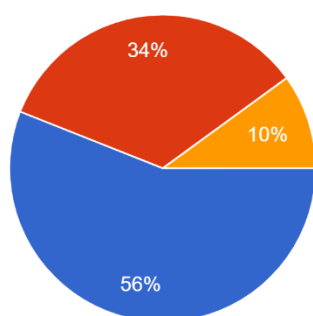
However, the highest humorous effect was achieved by a human-created joke: *'Газманова уволили с фабрики по производству календарей, потому что он оставлял ясные дни себе.'* The joke is culturally resonant, as it references a fragment of a well-known song performed by the

Russian singer Oleg Gazmanov: *'А я ясные дни оставляю себе, а я хмурые дни возвращаю судьбе'*. The joke is based on breaking of 'script oppositions' (our expectations), which leads to humorous results (solution type 2). Nevertheless, this translation version relies heavily on extralinguistic knowledge and may not be appreciated by respondents unfamiliar with the song. This joke was evaluated as humorous by 56% of respondents (see Figure 16).

**Figure 16.** Joke 7 human evaluation ratings for humorous effect

**Рисунок 16.** Оценка респондентами степени выраженности юмористического эффекта в шутке 7

Which joke do you think is the funniest?  
 150 responses



- Газманова уволили с фабрики по производству календарей, потому что он оставлял ясные дни себе.
- Невероятно, но факт: уволили меня с завода по производству календарей. Причина? Один пропущенный день.
- Не могу поверить, что меня уволили с фабрики календарей. Я всего лишь взял один день отгула.

Joke 7 textometric analysis demonstrates that human translators tend to prefer a more holistic approach and are not afraid of employing a different play element and/or imagery (see Table 13<sup>28</sup>). In case with Joke 7, AI-generated translations are more stereotypical and calque-oriented. Though these translations manage to maintain the same play element and imagery, the effectiveness of this solution does not prove its adequacy.

### 5.8.2. Joke 7 evaluation results

As shown in Figure 17, 72.3% of respondents correctly identified the third joke as AI-translated, whereas only 20.3% of respondents evaluated the second AI-generated joke as AI-generated. The human-translated version received the lowest percentage of votes attributing it to AI. The most frequently identified words and expressions that respondents believed indicated AI presence in the translation were: *'один день отгула'*, *'один пропущенный день'*, *'не могу поверить'*, *'фабрика календарей'*. Respondents also noted a general impression of high literality, which they associated with AI-generated translations without specifying particular words or

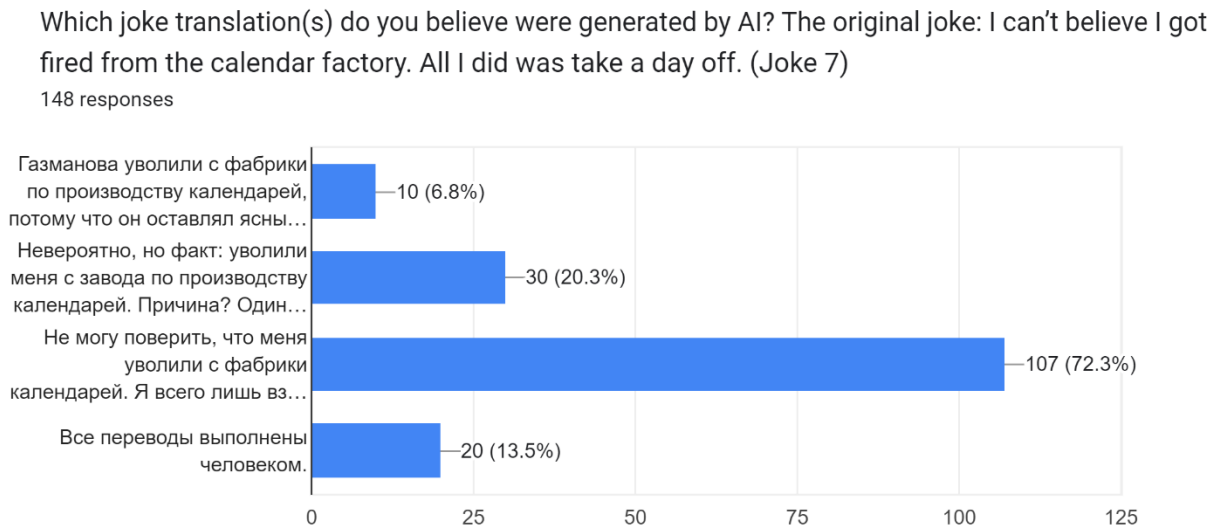
<sup>28</sup> <https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Tables/Table%2013.%20Linguistic%20features%20of%20Joke%207%20and%20its%20Russian%20translations.pdf>

phrases: literal translation, word-by-word translation. Human evaluation can be also

compared to automatic AI evaluation (see Table 14).

**Figure 17.** Respondents' identification of AI-generated translations in Joke 7

**Рисунок 17.** Обнаружение респондентами признаков ИИ в переводах шутки 7



**Table 14.** Automatic evaluation results for Joke 7

**Таблица 14.** Результаты автоматической оценки для шутки 7

AI Detector Tool	Translation 1 (HT)	Translation 2 (AI)	Translation 3 (AI)
AI Detector	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
BypassGPT	100% AI 0% HT	100% AI 0% HT	100% AI 0% HT
GrammarChecker	99% AI 1% HT	99% AI 1% HT	0% AI 100% HT

### 5.9. Data aggregation and discussion

Data aggregation makes it possible to formulate specific textometric tendencies that act as a set of criteria which can facilitate the identification of AI-generated and human-translated texts (Appendix 8<sup>29</sup>). One of the challenges for the academic society is to establish whether respondents representing the so-called general public are really able to perform such identification and measure not only translation equivalence, but also adequacy without any professional linguistics competences.

Following the logical framework of the joke-by-joke analysis above, this section summarizes and visualizes the overall parameters to denote any possible differences that create a gap between AI-generated and human-translated jokes.

Naturalness/artificiality analysis in the TEXT category shows that average Russian human translation results have a good correlation with the English source text (see Figure 18), with the number of sentences being slightly less and the sentence length being slightly bigger than in English. AI-generated translations demonstrate a 50% drop in the number of sentences and an obvious decrease in sentence length.

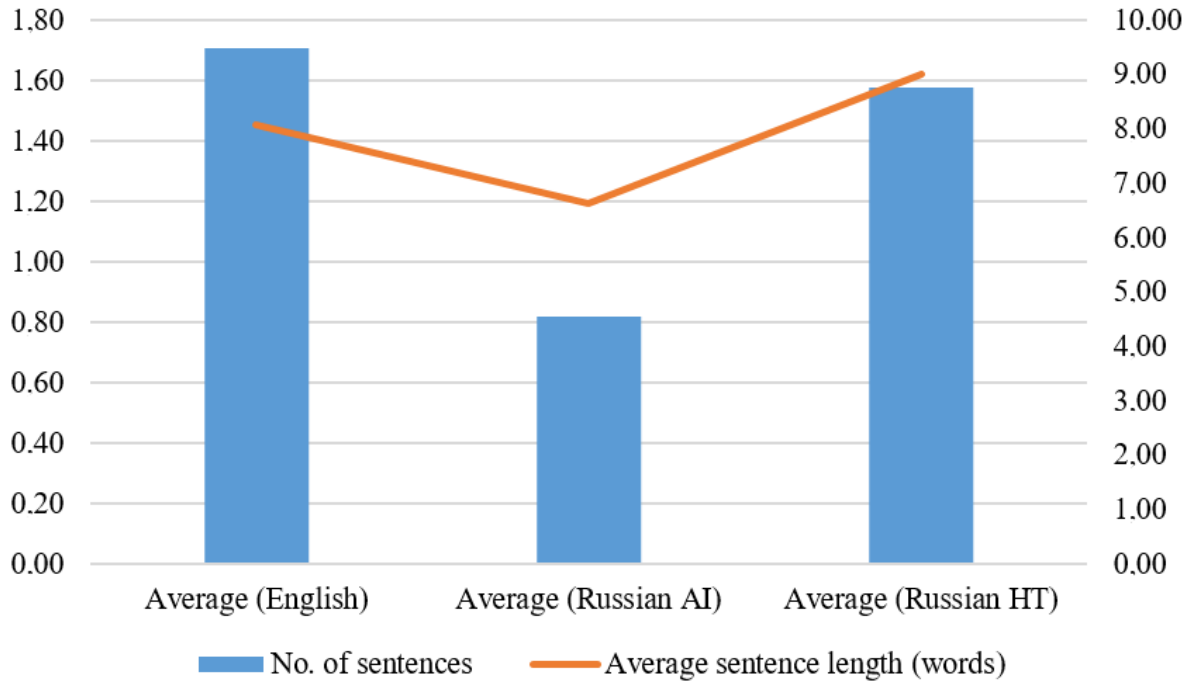
Such drastic differences usually involve increased structural complexity reflected by

<sup>29</sup><https://github.com/RudenkoElena88/Artificial-Vs-Human-Intelligence-in-Translation/blob/main/Appendixes/APPENDIX%208.pdf>

changes in punctuation, with fewer punctuation marks being used within a sentence and more expressive punctuation at the end of the sentence (see Figure 19).

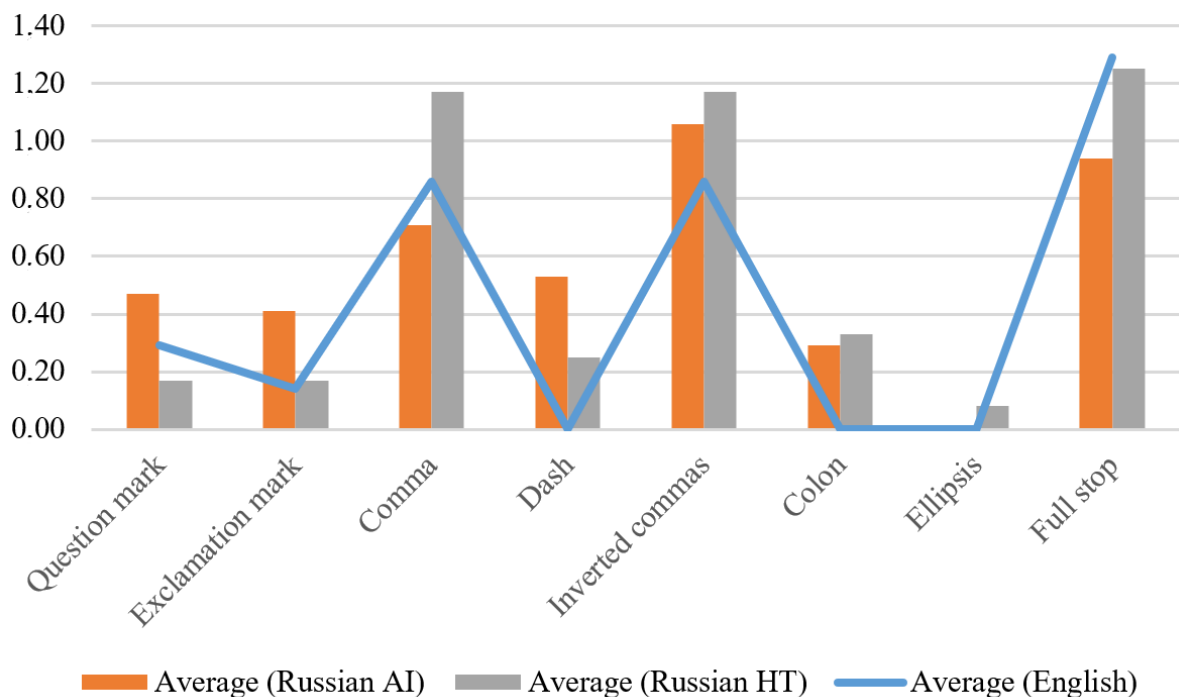
**Figure 18.** Analysis of naturalness confusion parameters for the TEXT category

**Рисунок 18.** Анализ параметров для определения естественности текста по категории ТЕКСТ



**Figure 19.** Analysis of naturalness confusion parameters for the PUNCTUATION category

**Рисунок 19.** Анализ параметров для определения естественности текста по категории ПУНКТУАЦИЯ

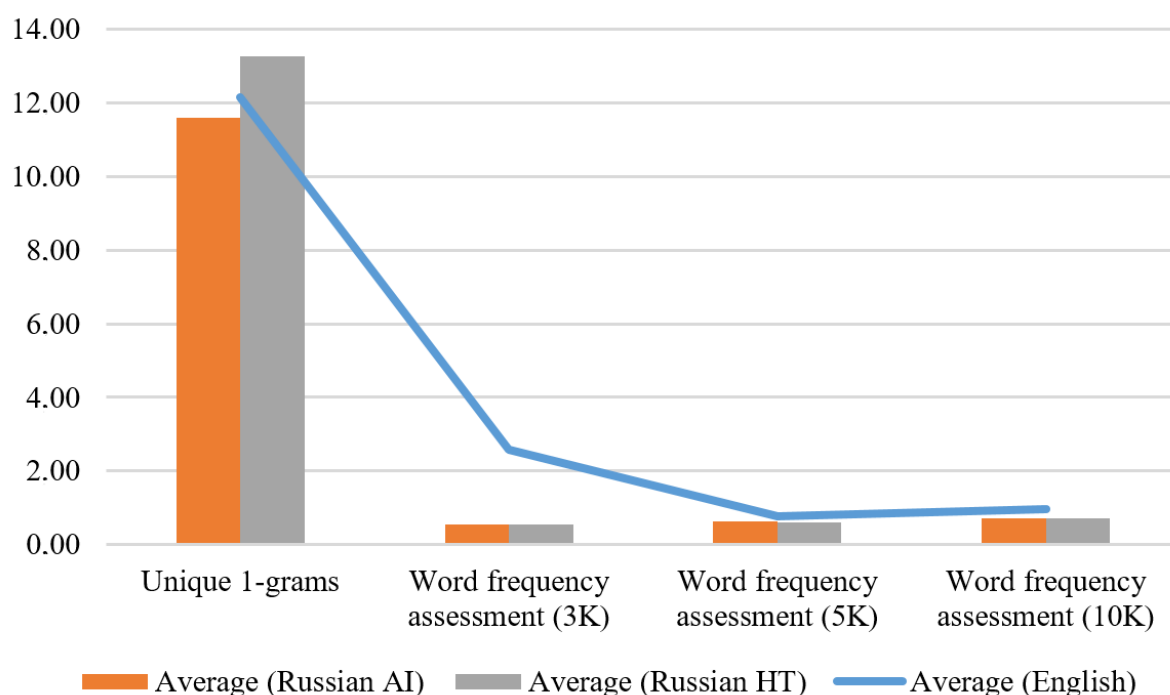


The repetitiveness category both in AI-generated and human translations quite closely follows the graph for the English source text, with a significant deviation only

for the 3K word frequency assessment, which features a more complex vocabulary used in Russian (see Figure 20).

**Figure 20.** Analysis of naturalness confusion parameters for the REPETITIVENESS category

**Рисунок 20.** Анализ параметров для определения естественности текста по категории ПОВТОРЯЕМОСТЬ

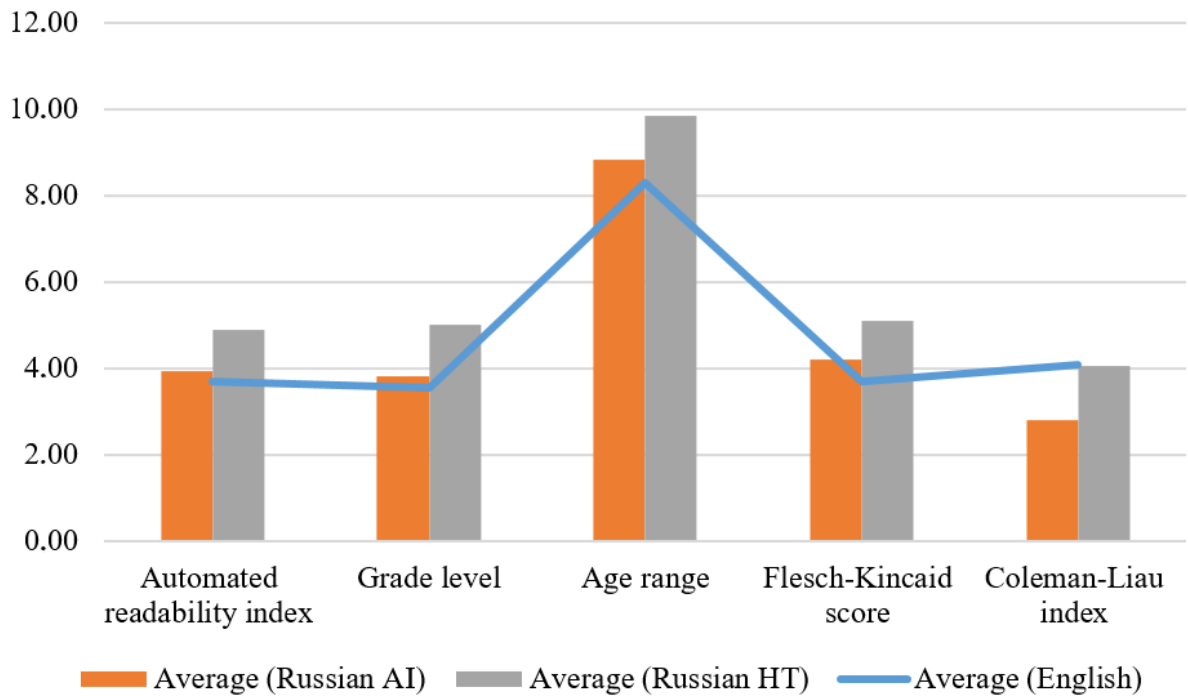


Readability analysis reveals almost a full match in scores between English source text and AI-generated translations (see Figure 21). Human translations into Russian feature higher scores than AI-generated texts for all index parameters under study. At the same time, it is important to note that the readability category did not have a significant impact on the respondents' reactions, which can be explained by the fact that all our respondents were university students and teaching staff while English jokes and their Russian translations are accessible for schoolchildren aged 8-10.

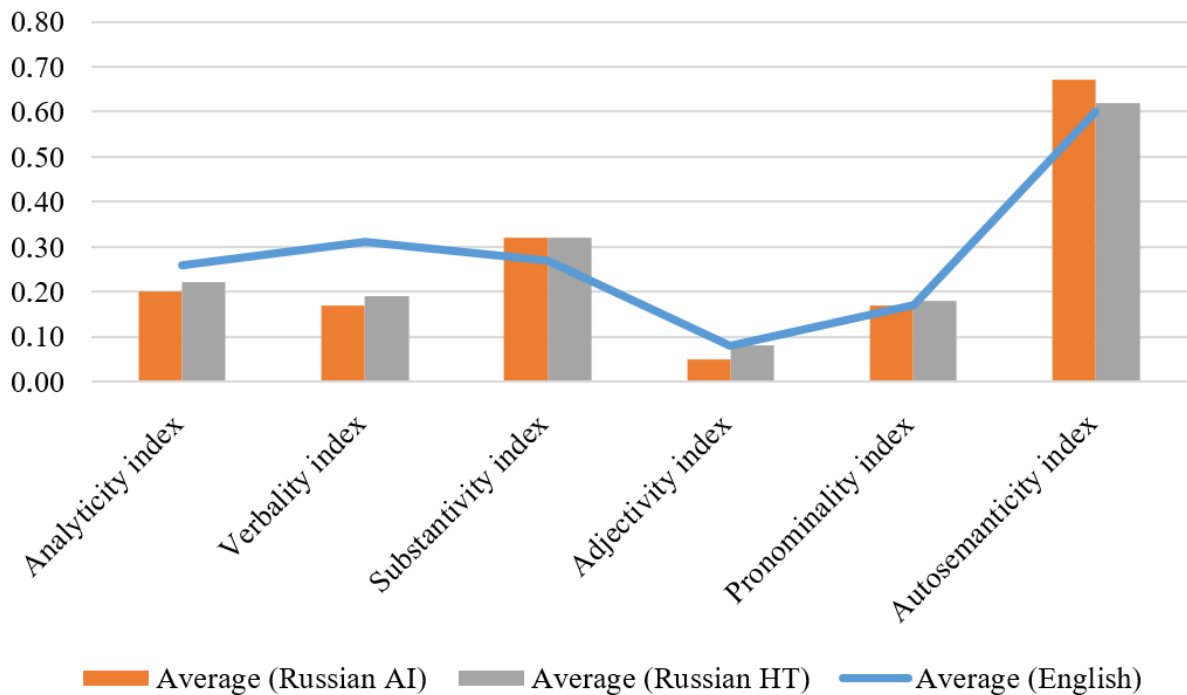
Part-of-speech aggregation shows that structurally all texts obtained are quite close to each other, which can be explained by systematic similarities between English and Russian, on the one hand, and by the genre of pun-based jokes selected for the present case study, on the other (see Figure 22). Such jokes feature stereotypical formal structure, which creates a serious limitation and sometimes can produce obstacles to achieve translation equivalence and adequacy. Following the general logical organization of such a joke, translators have applied a holistic approach to overcome this challenge.



**Figure 21.** Analysis of naturalness confusion parameters for the READABILITY category  
**Рисунок 21.** Анализ параметров для определения естественности текста по категории ЧИТАБЕЛЬНОСТЬ



**Figure 22.** Analysis of naturalness confusion parameters for the POS category  
**Рисунок 22.** Анализ параметров для определения естественности текста по категории ЧАСТИ РЕЧИ



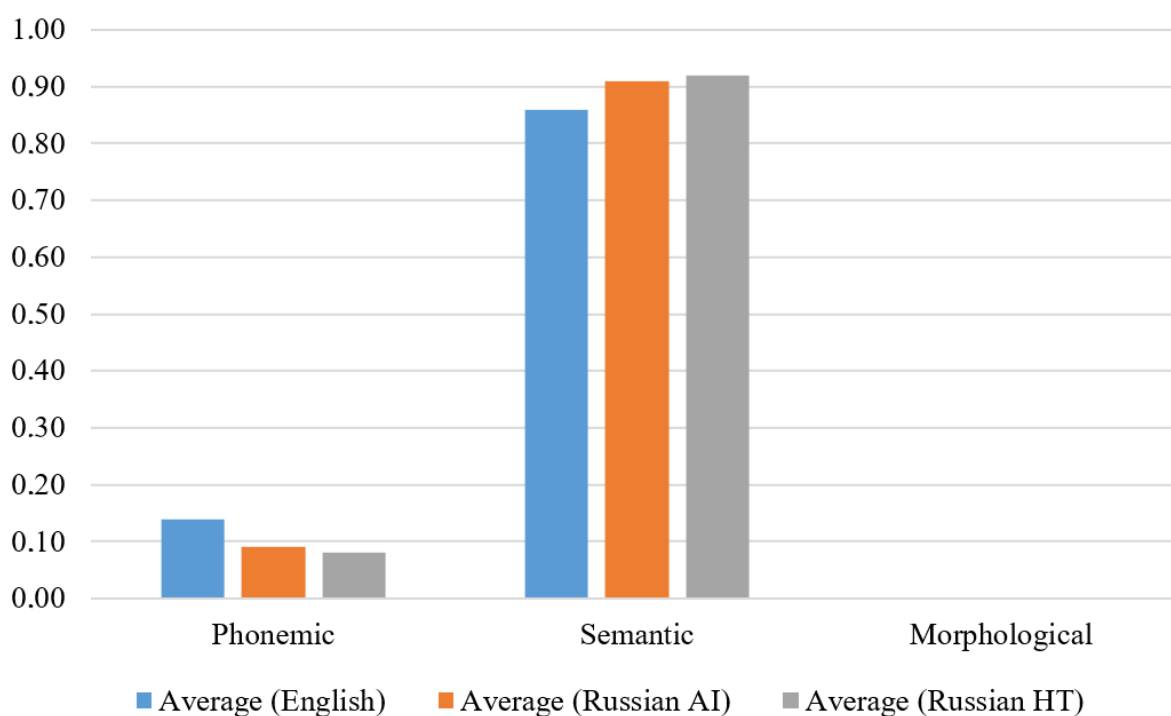
As far as humorous effect is concerned, not a single joke features a play element at the morphological level (see Figure 23). Absence of morphological-level play elements might explain why grammatical parameters and structural complexity, though evident in some translations, did not prevent our respondents to understand the jokes. In English, only one joke is based on a phonemic play element (Joke 1) while all other jokes contain semantic-level play elements. In Russian, neither AI-generated nor human-translated jokes include a pure phonemic play element. It can be assumed that the semantic level is fundamental for these Russian translations,

with certain phonemic components being sometimes used alongside with various semantic means.

There is a striking correlation in graphs for humorous effect rating and solution classification type for both Russian AI-generated and human-translated jokes (see Figure 24). It is obvious that the funnier a joke is for respondents, the more holistic its translation is, which means that a good translator of such texts should possess a greater creative freedom, avoid stereotypical play elements and/or imagery, and seek to generate novel interpretations.

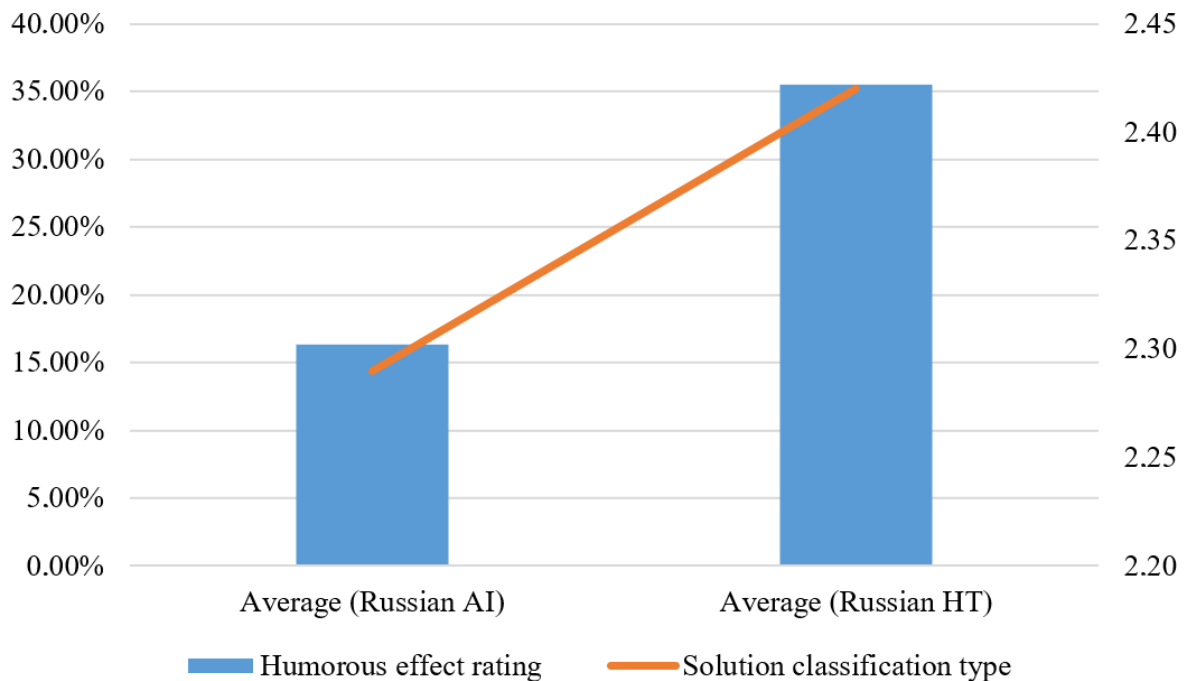
**Figure 23.** Analysis of naturalness confusion parameters for the HUMOUR EFFECT LEVEL category

**Рисунок 23.** Анализ параметров для определения естественности текста по категории УРОВЕНЬ ЮМОРИСТИЧЕСКОГО ЭФФЕКТА



**Figure 24.** Analysis of naturalness confusion parameters for the SOLUTION CLASSIFICATION TYPE and HUMOUR EFFECT RATING categories

**Рисунок 24.** Анализ параметров для определения естественности текста по категориям ТИП ПЕРЕВОДЧЕСКОГО РЕШЕНИЯ и СТЕПЕНЬ ВЫРАЖЕННОСТИ ЮМОРИСТИЧЕСКОГО ЭФФЕКТА



As our study shows, the translations that received the highest ratings for humorous effect were those produced by human translators (see Table 15). It should also be noted that AI-generated translations, which were evaluated as humorous within the range of 4.7% to 34%, were technically hybrid human-machine translations, as these AI-generated outputs often required further

revision. Nitzke et al. define revision competence as the ability to “develop strategies for consciously reading a text, not written by him” and to “handle the trade-off between the necessary changes and over-editing” (Nitzke et al., 2019: 249). In our case, this competence was divided between the human translator, who identified the issue, and AI-translator, which endeavored to fix it.

**Table 15.** Summary table of translations with the highest ratings for humorous effect

**Таблица 15.** Сводная таблица переводов с самой высокой степенью выраженности юмористического эффекта

№	The original joke	Human translation	Percentage
1	<i>What did the pirate say when he turned 80? Aye matey!</i>	<i>У пирата три тысячи друзей, вот только все они черти.</i>	51.3%
2	<i>Candy is dandy but liquor is quicker.</i>	<i>Конфеты – эффектны, но aperитивы – оперативней!</i>	34%
3	<i>I took the shell off my racing snail, thinking it would make him run faster. If anything, it made him more sluggish.</i>	<i>Улитка Ульяна ползла на работу так быстро, что выползла из домика.</i>	42%

4	<i>Two windmills are standing in a field and one asks, "What's your favorite kind of music?" The other says, "I'm a big metal fan."</i>	<i>Одна мельница говорит другой: «Как ты могла не поздравить меня с днём рождения?!» А та ей отвечает: «Извини, что-то совсем закрутилась».</i>	62%
5	<i>I have a split personality," said Tom, being frank.</i>	<i>– Кажется, у меня раздвоение личности, – сказал Том, глядясь в зеркало. – У нас, – мягко поправило отражение.</i>	58%
6	<i>I tried catching fog yesterday. Mist.</i>	<i>Я вчера гулял по Лондону, но мало что помню... Всё как в тумане.</i>	34%
7	<i>I can't believe I got fired from the calendar factory. All I did was take a day off.</i>	<i>Газманова уволили с фабрики по производству календарей, потому что он оставлял ясные дни себе.</i>	56%

The scores presented below reflect how well it is possible to identify AI-like elements in AI translations of pun-based jokes (see Table 16). The average percentage of respondents who correctly identified AI patterns in the AI-generated translations is approximately 55.6%. In case with automatic AI detectors, their scores are not consistent as such online tools available for checking Russian texts do not really perform this operation and are somehow preset to inform a user that the uploaded text is AI-generated. Thus, along with the low F1 Score for automatic evaluation, it can be assumed that AI detectors cannot be a reliable instrument and serve rather as an optional than an indispensable tool. Another issue that could possibly be a plausible explanation is that AI

detectors tend to become vulnerable the stronger LLMs become the more difficult it is to identified AI-generated texts (Zhang, Ma et al., 2024).

Confusion parameters calculation clearly states that human evaluation can be a reliable and indispensable tool in AI generation identification. F1 Score for human evaluation does not exceed the accepted level of evaluation consistency (80%) and does not match the Translation Quality Index (93.6%), but it is still far more reliable than automatic evaluations.

Finally, Table 17 enlists AI-like elements which the respondents used as their own set of criteria to identify AI-generated translations.

**Table 16.** Human vs automatic evaluation results

**Таблица 16.** Сравнительные результаты автоматической оценки и оценки, выполненной человеком

	<b>Human evaluation</b>	<b>Automatic evaluation</b>
TPR/Recall	0.7000	0.1133
FNR	0.3000	0.8667
TNR	0.5583	0.8933
FPR	0.4417	0.1067
Accuracy	0.6345	0.8893
Precision	0.6487	0.6145
F1 Score	0.6738	0.1913

**Table 17.** Respondents' comments on AI-generated content identification criteria

**Таблица 17.** Комментарии респондентов относительно критериев обнаружения признаков ИИ в переводах

Joke No.	Comments of respondents who correctly identified AI-generated translations (original spelling and punctuation preserved)
Joke 1	Начало в точности как в оригинале «Что сказал пират» «Семьдесят футов под килем!» – звучит как то, что мог бы написать только ИИ, несмешное объяснение Слишком логично
Joke 2	Наличие частиц «это»
Joke 3	гоночная улитка + конструкция предложения Я думаю, что ИИ оставляет большую часть слов. Слово «слизнявой» не передаёт суть шутки, потому что это не связано со скоростью. наличие слова «слизнявый», на мой взгляд, редкое в употреблении или такого слова не существует
Joke 4	кто вообще использует слово «ветряк» Контекст остается прежним. Я думаю, что ИИ не может представить шутку в другой интерпретации. отсутствует адаптация шутки, речевые «украшения», речь неживая
Joke 5	И здесь и в предыдущих шутках (да и дальше), ИИ пытается 1к1 перевести начало шуток «... будучи Франком» – в оригинале есть игра слов, а тут вообще ничего, слишком прямо
Joke 6	! в конце Сдувается странное слово Сдувается (нет такого выражения, что кто-то сдувается, только сливается)
Joke 7	Думаю, что нейросеть не смогла бы сопоставить «ясные дни» с песней Газманова Превалирует дословный перевод Неправильное построение предложения

## 6. Conclusions

As demonstrated in the Results, the survey was deliberately designed to first evaluate the joke's comic effect without revealing that these jokes were translated versions of the original English pun-based jokes, i.e., initially they were introduced as original Russian jokes. The aim was to measure respondents' text reception and to elicit their immediate reaction to several translation alternatives, devoid of any preliminary explanations that could have distorted their reception in the target linguistic and cultural system. In our study, the respondents' responses were taken as ultimate criteria of achieving joke's main purpose –

causing laughter and thereby generating pleasure. We align ourselves with the response-oriented approach to translation quality evaluation: “Übersetzungskritik <...> sollte immer klar diagnostizieren, welche Wirkung der übersetzte Text in seinem Umfeld und für seine Rezipienten hat.” (Translation criticism <...> should always clearly diagnose what effect the translated text has in its environment and for its recipients.) (Hönig, 2020: 123). There exists a contradiction between the two criteria of an adequate translation: the equipollency of source text (ST) and target text (TT) regulatory influences, and their semantic and structural similarity. According to Vermeer,

“in translation, priority has to be given to one factor and the others have to be subjected to it – because one cannot serve two masters at the same time” (Vermeer, 1994: 13). Therefore, we decided to prioritize the first criteria: the equipollency of ST and TT regulatory influences, as preserving the original semantics and syntax would have been impossible without losing the humorous effect. To empirically diagnose the effect of the translated text on recipients, we designed a survey that offered multiple-choice options with translations produced both by AI and a human.

The Results section apodictically demonstrates that the highest humorous response was achieved by human-translated jokes. In most cases, the human translator of pun-based jokes employed creative solutions of translation problems, such as partially changing or completely reformulating the joke *ex novo* in order to preserve its humorous potential (holistic translation technique).

The second part of the survey disclosed the nature of the jokes: respondents were informed that jokes were not original but translated, and that AI had participated in translating some of the jokes. Respondents were then asked to identify signs of AI involvement. In this part of the survey, there was no unanimity among respondents: on average, only 55.6% correctly identified AI patterns in the AI-generated translations, and 25.76% of respondents, on average, were convinced that all translations were produced by a human.

We hold the view that the first part of the survey clearly indicates humans’ ability to distinguish between AI- and human-created content on the basis of emotional decision-making. In contrast, the second part, which involved logical decision making, revealed respondents’ confusion and doubt, marked in their answers by the following words and expressions: ‘не знаю’, ‘не знаю, так кажется’, ‘чутье у меня’, ‘чувствуется’, ‘интуиция’, ‘затрудняюсь’.

Confusion parameters, which are defined by linguistic features of the dataset and provide an objective approach to identify and evaluate translation naturalness of pun-based jokes translated from English into Russian, can be classified into three groups:

1) nonconfusing parameters: number of sentences, average sentence length, POS (especially adjectives, pronouns and autosemanticity); solution classification type combined with humour effect rating;

2) neutral parameters: punctuation, repetitiveness, and humour effect level;

3) confusing parameters: readability.

To sum up, humans demonstrate more consistent evaluation of pun-based jokes translated from English into Russian than AI detection tools available for the Russian language. Hence humorous effect evaluation is valid only if performed by humans, whose emotional response turned out to be the only reliable metrics for assessing joke adequacy. AI can be recommended as an auxiliary tool for translating and assessing pun-based jokes, but cannot be regarded compulsory.

## References

- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y. and Fung P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 1, 675–718. <https://doi.org/10.18653/v1/2023.ijcnlp-main.45> (In English)
- Becker, J., Wahle, J. Ph., Gipp, B. and Ruas, T. (2024). Text generation: A systematic literature review of tasks, evaluation, and challenges, *arXiv:2405.15604v1*. <https://doi.org/10.48550/arXiv.2405.15604> (In English)
- Blinova, O. and Tarasov, N. (2022). A hybrid model of complexity estimation: Evidence from Russian legal texts, *Frontiers in Artificial Intelligence*, 5. <https://doi.org/10.3389/frai.2022.1008530> (In English)

- Çano, E. and Bojar, P. (2020). Human or machine: automating human likeliness evaluation of NLG Texts, *arXiv:2006.03189v1*. <https://doi.org/10.48550/arXiv.2006.03189> (In English)
- Celikyilmaz, A., Clark, E. and Gao, J. (2021). Evaluation of text generation: A survey, *arXiv preprint arXiv: 2006.14799*. <https://doi.org/10.48550/arXiv.2006.14799> (In English)
- Corizzo, R. and Leal-Arenas, S. (2024). One-GPT: A one-class deep fusion model for machine-generated text detection, *2023 IEEE International Conference on Big Data (BigData)*. <https://doi.org/10.1109/BigData59044.2023.10386674> (In English)
- Doughman, J., Afzal, O. M., Toyin, H. O., Shehata, Sh., Nakov, P. and Talat, Z. (2024). Exploring the limitations of detecting machine-generated text, *arXiv:2406.11073v1*. <https://doi.org/10.48550/arXiv.2406.11073> (In English)
- Etzaniz, J., Azkune, G., Soroa, A., Lopez de Lacalle, O. and Artetxe, M. (2023). Do multilingual language models think better in English?, *arXiv:2308.01223*. <https://doi.org/10.48550/arXiv.2308.01223> (In English)
- Fadaee, E. (2011). Translation naturalness in literary works: English to Persian, *Internal Journal of English and Literature*, 2 (9), 200–205. (In English)
- Fraser, K. C., Dawkins, H. and Kiritchenko, S. (2024). Detecting AI-generated text: factors influencing detectability with current methods, *arXiv:2406.15583v1*. <https://doi.org/10.48550/arXiv.2406.15583> (In English)
- Fu, J., Ng, S.-K., Jiang, Zh. and Liu, P. (2023). GSTScore: Evaluate as you desire, *arXiv preprint arXiv:2302.04166*. <https://doi.org/10.48550/arXiv.2302.04166> (In English)
- Gkatzia, D. and Mahamood, S. (2015). A snapshot of NLG evaluation practices 2005–2014, *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, Brighton, September 2015, 57–60, *arXiv:2302.07459v2*. <https://doi.org/10.18653/v1/W15-4708> (In English)
- Goddard, J., Celik, Y., Goel, S. (2024). Beyond the human eye: Comprehensive approaches to AI text detection, *18<sup>th</sup> Annual Symposium on Information Assurance (ASIA '23)*, 53. (In English)
- Gryka, P., Gradoń, K., Kozłowski, M., Kutyla, V. and Janicki, A. (2024). Detection of AI-Generated Emails – A Case Study, *The 19<sup>th</sup> International Conference on Availability, Reliability and Security (ARES 2024)*, July 30 – August 02, 2024, Vienna, Austria. ACM, New York, USA. <https://doi.org/10.1145/3664476.3670465> (In English)
- He, Zh., Liang, T., Jiao, W., Zhang, Zh., Yang, Y., Wang, R., Tu, Zh., Shi, Sh. and Wang, X. (2023). Exploring humanlike translation strategy with large language models, *arXiv:2305.04118*. <https://doi.org/10.48550/arXiv.2305.04118> (In English)
- Hendy, A., Abdelrehim, M. G., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M. and Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation, *arXiv.2302.09210*. <https://doi.org/10.48550/arXiv.2302.09210> (In English)
- Hönig, H. G. (2020). *Konstruktives Übersetzen*, Stauffenburg Verlag, 3rd ed. (In German)
- Jiao, W., Wang, W., Huang, J. T., Wang, X., Shi, Sh. and Tu, Zh. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine, *arXiv:2301.08745*. <https://doi.org/10.48550/arXiv.2301.08745> (In English)
- Li, Ch., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q. and Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli, *arXiv:2307.11760*. <https://doi.org/10.48550/arXiv.2307.11760> (In English)
- Mossop, B. (2000). The workplace procedures of professional translators, in Chesterman, A., San Salvador, N. G., and Gambier, Y. (eds.), *Translation in context: selected papers from the EST Congress*, Granada 1998, John Benjamins, Amsterdam, 39–48. <https://doi.org/10.1075/btl.39.07mos> (In English)
- Nitzke, J., Hansen-Schirra, S. and Canfora, C. (2019). Risk management and post-editing competence, *Journal of Specialised Translation*, 31, 239–259. (In English)

- Obeidat, A. M., Ayyad, G. R., Sepora, T. and Mahadi, T. (2020). The tension between naturalness and accuracy in translating lexical collocations in literary text, *Journal of Social Sciences and Humanities*, 17 (8), 123–134. (In English)
- Oedingen, M., Engelhardt, R. C., Denz, R., Hammer, M., Konen, W. (2024). ChatGPT code detection: Techniques for uncovering the source of code, *AI 2024*, 5, 1066–1094. <https://doi.org/10.3390/ai5030053> (In English)
- Pan, W. H., Chok, M. J., Shan Wong, J. L., Shin, Y. X., Poon, Y. Sh., Yang, Zh., Chong, Ch. Y., Lo, D. and Lim, M. K. (2017). Assessing AI detectors in identifying AI-generated code: Implications for education, *Conference'17*, Washington, DC, USA, *arXiv:2401.03676v1*. <https://doi.org/10.48550/arXiv.2401.03676> (In English)
- Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English, *Language*, 55 (2), 291–318. <https://doi.org/10.2307/412586> (In English)
- Puduppully, R., Kunchukuttan, A., Dabre, R., Aw, A. T. and Chen, N. F. (2023). Decomposed prompting for machine translation between related languages using large language models, *arXiv:2305.13085*. <https://doi.org/10.48550/arXiv.2305.13085> (In English)
- Rogers, M. (1999). Naturalness and translation, in: Rasmussen, W., Roald, J. and Simonnæs, I. (eds). *SYNAPS 2*. Bergen: NHH, 9–31. (In English)
- Schuff, H., Vanderlyn, L., Adel, H. and Vu, Th. (2023). How to do human evaluation: A brief introduction to user studies in NLP, *Natural Language Engineering*, 29, 1–24. <https://doi.org/10.1017/S1351324922000535> (In English)
- Sellam, Th., Das, D. and Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation, *arXiv:2004.04696*. <https://doi.org/10.48550/arXiv.2004.04696> (In English)
- Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D. and Wei, J. (2023). Language models are multilingual chain-of-thought reasoners, *The Eleventh International Conference on Learning Representations*, *arXiv:2210.03057*. <https://doi.org/10.48550/arXiv.2210.03057> (In English)
- Vermeer, H. J. (1994). Translation today: Old and new problems, in Snell-Hornby, M., Pöchhacker, F. and Kaindl, K. (eds.), *Translation Studies – An Interdiscipline*, John Benjamins, Amsterdam/Philadelphia, 3–16. (In English)
- Wang, Z. M., Peng, Zh., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Zhang, M., Zhang, Zh., Ouyang, W., Xu, K., Chen, W., Fu, J. and Peng, J. (2023). Role LLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models, *arXiv.2310.00746*. <https://doi.org/10.48550/arXiv.2310.00746> (In English)
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V. and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 35, 24824–24837. <https://doi.org/10.48550/arXiv.2201.11903> (In English)
- Yang, Ch., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D. and Chen, X. (2023). Large language models as optimizers, *arXiv:2309.03409*. <https://doi.org/10.48550/arXiv.2309.03409> (In English)
- Yang, X., Zhan, R., Wong, D. F., Wu, J. and Chao, L. S. (2023). Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX*, Macau SAR, China, Machine Translation Summit, 2, 88–98. <https://doi.org/10.48550/arXiv.2310.08908> (In English)
- Zabalbeascoa, P. (2005). Humor and translation – an interdiscipline, available at: <https://www.semanticscholar.org/paper/Humor-and-translation-an-interdiscipline-Zabalbeascoa/2cdcc20591c39f32bd5e2049534600515ede3f3f> (Accessed 25.06.2024). <https://doi.org/10.1515/humr.2005.18.2.185> (In English)
- Zhang, Y., Ma, Y., Liu, J., Liu, X., Wang, X. and Lu, W. (2024). Detection Vs. Anti-detection: Is text generated by AI detectable?, *iConference 2024*, LNCS 14596, 209–222. [https://doi.org/10.1007/978-3-031-57850-2\\_16](https://doi.org/10.1007/978-3-031-57850-2_16) (In English)
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2020). BERTscore: Evaluating text generation with



BERT, *arXiv preprint arXiv: 1904.09675*.  
<https://doi.org/10.48550/arXiv.1904.09675> (In English)

*All authors have read and approved the final manuscript.*

*Все авторы прочитали и одобрили окончательный вариант рукописи.*

*Conflicts of interests: the authors have no conflicts of interest to declare.*

*Конфликты интересов: у авторов нет конфликтов интересов для декларации.*

**Elena S. Rudenko**, Candidate of Philology, Associate Professor, Associate Professor of the Department of Integrative and Digital Linguistics, Don State Technical University, Rostov-on-Don, Russia.

**Руденко Елена Сергеевна**, кандидат филологических наук, доцент, доцент кафедры «Интегративная и цифровая лингвистика»,

Донской государственный технический университет, Ростов-на-Дону, Россия.

**Marina Yu. Semenova**, Candidate of Philology, Associate Professor, Head of the Department of Integrative and Digital Linguistics, Don State Technical University, Rostov-on-Don, Russia.

**Семенова Марина Юрьевна**, кандидат филологических наук, доцент заведующий кафедрой «Интегративная и цифровая лингвистика», Донской государственный технический университет, Ростов-на-Дону, Россия.

#### **CRedit author statement**

**Rudenko Elena Sergeevna:** conceptualization, resources, investigation, software, validation, writing – original draft; **Semenova Marina Yurievna:** data curation, methodology, formal analysis, project administration, visualization, supervision, writing – review and editing.