Automated Speech Act Annotation in a Russian Spoken Corpus Using Large Language Models: A Comparative Study

Tatiana Sherstinova, Viktoria Firsanova, Alena Novoseltseva, Mariya Megre, Egor Savchenko
National Research University Higher School of Economics
Saint Petersburg, Russia
{tsherstinova, vfirsanova}@hse.ru, {aonovoseltseva, msmegre, easavchenko_1}@edu.hse.ru

***Abstract*—The research focuses on the automatic annotation of a linguistic corpus using large language models (LLMs). Annotating a corpus is a crucial step in its creation, as it determines the practical scope and applications of the resource being developed. This study explores the annotation of oral speech transcripts at the pragmatic level using speech acts that reflect the speaker's intent and purpose. Typically, this task is performed manually by experts, which greatly limits the volume of annotated data that can be produced. In this work, an attempt was made to automatically annotate speech acts using five LLMs commonly used for processing Russian texts – ChatGPT, GigaCHAT, YandexGPT, Mistral, and Gemini. A comparative analysis of the automatic annotation results was conducted, highlighting the strengths and weaknesses of each model. . The findings suggest that employing LLMs for corpus annotation is a promising approach, with ChatGPT and Gemini demonstrating particular effectiveness in speech act categorization. However, for Russian, language-specific models like GigaCHAT and YandexGPT are preferred when language-specific information is needed.**

I. Introduction

The research is dedicated to solving an important practical task: automating the annotation of linguistic corpora. A linguistic corpus is an electronic language resource used for scientific research, statistical processing of linguistic phenomena, and training language models on textual materials from these corpora.

Corpus annotation is a crucial step in the creation of a linguistic resource, as it enables searching through annotated linguistic data categories, filtering linguistic material, and obtaining statistical data on the conditions of annotated units' implementation. Each linguistic resource is characterized by its own set of annotation levels. The most standard ones include 1) the morphological level, where POS tagging (noun, verb, adjective, etc.), word forms (gender, number, case), and morphemes (roots, prefixes, suffixes) are annotated; 2) the syntactic level, describing syntactic connections between words in a sentence and syntactic constructions; and 3) the semantic level, where word and phrase meanings, semantic roles, and ontological categories (entity, event, property) are reflected. Annotation at these levels is usually done semi-automatically—first, appropriate parsers are used, and then expert review is conducted on the automatically generated annotations.

The material for this study was not written texts, as is usually the case, but transcriptions of audio recordings of everyday conversations in Russian from the ORD corpus, known as the "One Day of Speech" corpus [Asinovsky et al. 2009], [Bogdanova-Beglarian et al. 2016]. These texts reflect real, unscripted verbal communication on both personal and professional topics between two or more speakers in natural settings (at home, at work, in a store, in an office). The transcriptions reflect all the “imperfections” of spontaneous spoken language, both at the lexical and grammatical levels, which significantly complicates their annotation.

Furthermore, the task of this research is to annotate the transcriptions of sound recordings at the pragmatic level [Jurafsky 2006; Weisser 2014]—more specifically, at the level of speech acts that differ in the pragmatic goal that the speaker sets for themselves when producing an utterance (e.g., “statement”, “request”, “question”, etc.). Essentially, this involves a multidimensional classification of a representative set of phrases into various types of speech acts using the scheme developed for the ORD corpus (see Section II for details). This task was previously carried out exclusively by experts through manual annotation [Sherstinova 2016].

The emergence of large language models (LLMs) has raised the question of whether their capabilities can be applied to the automatic classification of utterances (phrases) into speech acts—that is, into different groups based on the speaker's pragmatic intent when producing the phrase. The primary goal of this research is to provide an answer to this question.

The article is structured as follows. Section II introduces the principles of annotating the ORD corpus by speech acts and the features of its notation. Section III describes the application of LLMs for solving similar tasks and describes related works. Section IV presents five experiments on the automatic identification of speech acts using different LLMs most commonly used for processing Russian-language data. Finally, in Section V, we discuss and interpret the obtained results.

II. Principles of speech act annotation in the ORD corpus

The concept of a speech act is currently quite widespread in modern linguistics, yet the common understanding of what constitutes a speech act and which categories are distinguished varies significantly across different linguistic traditions and schools. Initially, the term was introduced by J.L. Austin [Austin 1962], after which it underwent significant revision by J.R. Searle [Searle 1976], and subsequently, speech act theory was developed by many of their followers. Notable scholars who have contributed to the understanding of speech acts include A. Wierzbicka [Wierzbicka 1973], M.M.Bakhtin [Bakhtin 1986], and Yu.D. Apresyan [Apresyan 1986].

In pragmatic studies conducted on data of the ORD corpus, a speech act is understood as a purposeful speech action, considered within the context of a pragmatic situation and possessing a certain illocutionary force [Sherstinova 2015]. It is assumed that every speaker's utterance consists of one or more speech acts. Furthermore, we believe that any statement can be interpreted as a speech act of a certain type. In this sense, our approach differs from the traditional understanding of speech acts proposed by J.R. Searle.

In developing a classification scheme for speech acts for the annotation of the ORD corpus, we analyzed the most well-known systems for the formal representation of speech actions used in various linguistic corpora with pragmatic annotation, such as the SPAACy (Speech Act Annotated Corpus, UK) [Weisser 2003]; [Leech & Weisser 2003], the Dialogue Act Markup in Several Layers (DAMSL) system [Allen & Core 1997], the international Cross-Cultural Study of Speech Act Realization Patterns project [Blum-Kulka & Olshtain 1984], the Verbal Response Modes (VRM) discourse taxonomy system proposed by W. Stiles [Stiles 1992], and others.

However, most of the proposed classifications were developed for a limited set of communicative scenarios (e.g., phone calls to call centers or purchasing train tickets) and therefore are not suitable for annotating such a complex genre as everyday spoken communication. To address our task, it seemed appropriate to use speech act classifications developed by Russian linguists specifically for Russian conversational speech [Borisova 2009].

The main types of speech acts annotated in the ORD corpus are defined as follows [Sherstinova 2016]:

1. Representatives are speech acts whose primary goal is the exchange of information between dialogue participants.

2. Directives are speech acts intended to prompt the addressee to action (or inaction) or express an attempt to influence their worldview, emotions, and attitudes.

3. Commissives involve the speaker taking on certain commitments.

4. Expressives-emotives are used to express and convey feelings and emotions.

5. Etiquette expressives are standardized forms that regulate communication in polite and ritualized situations.

6. Valuatives are used to express evaluative opinions or judgments.

7. Suppositives express the speaker's opinion or assumption.

8. Communicative regulatives are phatic speech acts related to the "organizational" aspects of interaction, used to structure and manage dialogue.

In addition to these common classes of speech acts, in ORD pragmatic annotation the following other categories are used:

9. Undefined is used in cases (most often for incomplete speech fragments), where the illocutionary force cannot be determined.

10. Paralinguistic speech events, many of which may carry illocutionary force (e.g., laughter, sighs, groans, etc.).

For the purpose of this study, it was also decided to separate “questions” which were previously categorized during corpus annotation as a sub-type of representatives into their own group. This led to a new category:

11. Rogatives used to denote questions.

Within each major type of speech act, subtypes are distinguished because it may be necessary to separate, for example, a request from a command within the general category of directives [Sherstinova 2018]. However, in this particular study, we do not consider speech act subtypes, focusing only on the main categories.

When annotating speech acts in the ORD corpus, each utterance is listened to, segmented into fragments homogeneous in illocutionary force, and each fragment is assigned the closest corresponding speech act or a combination of speech acts when the same speech fragment performs multiple illocutionary functions simultaneously. To date, more than 200 communicative macro-episodes have been annotated in this way. The statistical distribution of speech acts within this material is presented in the article [Sherstinova et al. 2022].

When working with textual data, the notation—the principles of representing spoken speech in the corpus transcripts—is also of great importance. The ORD corpus uses a complex system of special symbols that experts are trained to understand, and these must obviously be included in the training instructions for the language model. Detailed information on the transcription notation used in the ORD corpus is presented in the following publications [Asinovsky et al. 2009; Sherstinova et al. 2010].

III. Large Language Models for Corpus Annotation: Related Works

Large Language Models (LLMs) are deep learning models for natural language generation trained on vast amounts of text data. LLMs use the Transformer architecture introduced in [Vaswani et al. 2017], which uses a self-attention mechanism to capture deep contextual relationships between words. LLMs shifted the paradigm of natural language processing from task-specific models to foundation models [Bommasani et al. 2021] that can solve a wide range of tasks. The foundation models can be guided toward the desired output through prompt engineering [Marvin et al. 2023], i.e. a process of building instructions for generative models.

The study focuses on applying the following state-of-the-art LLMs to the linguistic corpora annotation: ChatGPT [Achiam et al. 2023], GigaCHAT [GigaChat 2024], YandexGPT [YandexGPT 3 2024], Mistral [Jiang et al. 2023], and Gemini [Team G. et al. 2023]. The research focuses on the following LLM features:

1) capturing linguistic labels from user instructions,
2) refining outputs based on user feedback,
3) handling large context windows, processing documents,
4) processing data in Russian.

Meta-learning is the ability of a machine learning model to solve new tasks without being explicitly trained on them [Schmidhuber 1987]. For example, GPT-3 [Brown et al. 2020] and later GPT generations have few-shot capabilities that allow the LLM to label large amounts of data given a few examples of labeled samples. In corpus annotation, the LLMs are expected to capture linguistic categories from a few labeled samples in user prompts. The study assesses the LLMs' capacity to capture speech act descriptions based on labeled samples provided in the user instructions.

Iterative refinement is the ability of an LLM to correct its output based on user feedback [Chen et al. 2023]. In corpus annotation, refining outputs based on the follow-up prompts are necessary to align the model judgments with annotator decisions to ensure that automated annotation does not contradict linguistic information represented in the corpus.

The context window size of an LLM refers to the number of tokens (e.g., subwords) a model can process. For example, the context window size of state-of-the-art LLMs may vary from 8k to 128k tokens or more. Handling large context windows allows for processing documents, which is beneficial for automated corpora annotation. For example, document handling and large context window size allow for uploading partly annotated tabular data and returning a fully annotated table. In this example, the partial annotation plays the role of a few-shot sample set.

One of the challenges in applying LLMs to linguistic tasks is their multilingualism. The study highlights processing data in Russian. While ChatGPT, Gemini, and Mistral can show multilingual capacity, they were not explicitly trained to process Russian. In this study, the models are compared to GigaCHAT and YandexGPT expressly trained by Russian data.

The following sections describe the experiments applying ChatGPT, GigaCHAT, YandexGPT, Mistral, and Gemini to speech act corpus annotation. The experiments allow for assessing the LLMs' capacity toward capturing linguistic labels from user prompts, correcting outputs based on user-machine interaction history, handling large documents, and processing data in Russian.

IV. Experiments in Automatic Speech Act Annotation with Large Language Models

To evaluate the effectiveness of applying LLM to speech act annotation, experiments were conducted using the most commonly used language models when working with the Russian language, such as ChatGPT, GigaCHAT, YandexGPT, Mistral, and Gemini. Textual data used for the experiments was a subcorpus of the ORD corpus annotated at the speech act level. The research sample consists of 42,387 speech acts (phrases) related to 243 speech episodes. The results of the annotation were exported from the database as Excel tables containing the following fields: phrase number (#), unique episode code (*Episode*), start time of the utterance from the beginning of the file (*Time*), orthographic record of the phrase using ORD notation (*Phrase*), speaker code (*Speaker*), type of speech act (*SpeechAct*), subtype of speech act (*SubType*), and the speaker's social role in the given communicative situation (*SocRole*) (see Fig. 1).

#	Eposide	Time	Phrase	Speaker	SpeechAct	SubType	SocRole
1	ordS001-01	00:00:09.046	ещё пожалуйста сигареты / Vogue$ //	S001	ИНФ	ЗАП	КЛ
2	ordS001-01	00:00:12.844	а ?	S001	РЕГ	ПЕРЕ	КЛ
3	ordS001-01	00:00:13.298	какой Vogue$?	FN001	ИНФ	ВОПР	СС
4	ordS001-01	00:00:13.972	с ментолом //	S001	ИНФ	ОТВ	КЛ
6	ordS001-01	00:00:16.417	мне / взять ?	S001	ИНФ	ВОПР	КЛ
7	ordS001-01	00:00:18.020	возьми //	S002	ИНФ	ОТВ	КЛ
8	ordS001-01	00:00:18.597	это вместе ? да ?	FN001	ИНФ	ВОПР	СС
9	ordS001-01	00:00:19.729	не / вместе / вместе //	S002	ИНФ	ОТВ	КЛ
10	ordS001-01	00:00:20.453	вместе ?	FN001	РЕГ	ПЕРЕ	СС
11	ordS001-01	00:00:20.868	потом рассчитаемся //	S002	КОМ	ЗАЯН	КЛ
13	ordS001-01	00:00:23.920	что ещё ?	FN001	ИНФ	ВОПР	СС
14	ordS001-01	00:00:24.678	всё //	S002	ИНФ	ОТВ	КЛ

Fig. 1. Fragment of a Table with Exported Utterances Annotated at the Speech Act Level

These data were used for further work with LLMs. If the interaction with the LLM allowed for the entire file to be processed (e.g., for ChatGPT), the file was uploaded in full; otherwise, subsets of utterances were uploaded directly into the model's chat prompt. Communication with all language models was conducted in Russian.

A general communication protocol was developed for the experiments to simplify the comparison of the results. However, it was not always possible to fully adhere to it due to the specific features and interfaces of each model. The communication protocol generally proceeded as follows:

Step 1. Introducing the model to the corpus notation: loading instructions on how to interpret the transcription of utterances and what non-standard symbols are used in the transcripts. Checking that the model correctly understood the instructions.

Step 2. Introducing the model to the customized classifycation of speech acts with examples similar to the following (see Fig. 2).

Step 3. Requesting the model to formulate a detailed description of 2-5 sentences for each type of speech act based on the provided examples. Evaluating the results and refining the descriptions until a satisfactory scheme was achieved.

Step 4a. Tasking the model to determine the type of speech act for a list of given utterances.

Step 4b. Tasking the model to determine the type of speech act for the entire uploaded table (in csv or MS Excel format). In this case, additional information was provided to the model about the table structure and which columns needed to be analyzed, while columns with existing annotations were to be ignored.

SPEECH ACT TYPE phrase example

КОМИССИВЫ потом рассчитаемся //, ну-ка / давайте глянем //

ЭКСПРЕССИВЫ-ЭТИКЕТНЫЕ # очень (...) приятно /, здравствуйте //

ДИРЕКТИВЫ позвоните мне вечерком //, да / свою забирайте //

РЕГУЛЯТИВЫ КОММУНИКАТИВНЫЕ ну вот //, хорошо ?

РЕПРЕЗЕНТАТИВЫ не / вместе / вместе //, пятый нормальный //

РОГАТИВЫ это у шестого / да ?, а с пятым всё нормально ?

ВЕРДИКТИВЫ-ВАЛЮАТИВЫ она старая //, что-то с ним не то //

ФРАГМЕНТ НЕОПРЕДЕЛЯЕМЫЙ на(:) ..., а вот за(:) ... за вот з... шестна... вот вот эт...

ПАРАЛИНГВИСТИКА *С, *В

Fig. 2. The list of speech acts with examples for training models

Step 5. Evaluating the results of the automatically generated annotations, adjusting prompts to achieve the best outcome, and repeating steps 4 and 5 until the desired result was reached.

A. ChatGPT

The model used in this experiment was ChatGPT version 4.0.

Step 1. The model comprehended the transcript notation quite well and was able to provide examples for each element. Besides, it could calculate the statistics of symbols usage in the table data.

Step 3. The model successfully generated descriptions for each of the proposed types of speech acts. In cases where it provided inaccurate results, significant improvements were achievable through further chat.

1. **КОМИССИВЫ**: Речевые акты, в которых говорящий берет на себя обязательства выполнить действие в будущем, например "потом рассчитаемся / ну-ка / давайте глянем". Здесь выражаются обещания или намерения.
2. **ЭКСПРЕССИВЫ-ЭТИКЕТНЫЕ**: Включают вежливые формулы и выражения социального этикета, такие как приветствия или прощания, например "очень (...) приятно / здравствуйте".
3. **ДИРЕКТИВЫ**: Эти акты включают команды, просьбы или рекомендации, направленные на побуждение слушателя к действию. Пример: "позвоните мне вечерком /, да / свою забирайте".
4. **РЕГУЛЯТИВЫ КОММУНИКАТИВНЫЕ**: Речевые акты, направленные на управление процессом коммуникации, упорядочивание диалога или уточнение деталей, как в примере "я вам не сделаю не шестнадцать / а восемнадцать сделаю / хорошо?".
5. **РЕПРЕЗЕНТАТИВЫ**: Утверждения или высказывания, выражающие мнение или факты, которые говорящий считает верными, как в примере "пятый нормальный / мне кажется / что пятый / ...".

Step 4a. The model successfully handled the annotation of the list of phrases through prompts (See Fig. 3).

Step 4b. The model managed to annotate the entire large table (see Fig. 1), containing more than 40,000 phrases, by adding a new column with speech act annotations. However, the results were less convincing here. In the initial part of the table, the annotations were of high quality, but the accuracy significantly decreased later. In Tab. I, the percentage of speech acts identified by ChatGPT is compared with expert annotations for the same data. It is noticeable that a substantial portion of the phrases were classified as UNDEFINED, indicating the model's difficulty in identifying speech act types. However, it's important to note that these results were obtained without performing Step 1 (learning the notation).

Для классификации речевых актов я проанализировал каждую реплику:

1. **ещё пожалуйста сигареты / Vogue$ // — ДИРЕКТИВЫ** (просьба или заказ чего-либо).
2. **а ? — РОГАТИВЫ** (вопрос или уточнение).
3. **какой Vogue$? — РОГАТИВЫ** (вопрос для уточнения детали).
4. **с ментолом // — РЕПРЕЗЕНТАТИВЫ** (утверждение, уточнение деталей заказа).
5. ***П — ПАРАЛИНГВИСТИКА** (невербальные элементы, пауза или звук).
6. **мне / взять ? — РОГАТИВЫ** (вопрос или предложение с просьбой подтвердить действие).
7. **возьми // — ДИРЕКТИВЫ** (приказ или рекомендация).
8. **это вместе ? да ? — РОГАТИВЫ** (вопрос на уточнение состояния или объединения).
9. **не / вместе / вместе // — РЕПРЕЗЕНТАТИВЫ** (утверждение с возможным исправлением).

Fig. 3. The results of Step 3 for chatGPT

TABLE I. DISTRIBUTION OF SPEECH ACT TYPES BASED ON EXPERT ANNOTATION AND THAT PERFORMED BY CHATGPT (INITIAL AND FINE-TUNED)

Speech acts	Expert annotation		ChatGPT (initail)		ChatGPT (fine-tuned)	
	Sum	%	Sum	%	Sum	%
REPRESENTATIVES	18210	43,13	633	1,33	22823 ↑	50,48
REGULATIVES	7045	16,69	8465	17,80	9 ↓	0,02
ROGATIVES	5592	13,24	0	0	13820 ↑	30,57
DIRECTIVES	3262	7,73	330	0,69	246 ↓	0,54
VALUATIVES	3065	7,26	0	0	26 ↓	0,06
ETIQUETTE EMOT.	1427	3,38	0	0	911	2,01
SUPPOSITIVES	1229	2,91	0	0	281 ↓	0,62
COMMISSIVES	859	2,03	1	0,00	124 ↓	0,27
EMOTIVES	701	1,66	0	0	560	1,24
PARALINGUISTICS	485	1,15	5329	11,21	2510 ↑	5,55
UNDEFINED	348	0,82	32801	68,97	3904 ↑	8,63

Subsequent experiments aimed at improving the recognition of speech acts, both for specific categories (to teach the model to identify all etiquette forms, for instance) and using multi-class classification. The primary goal was to reduce the number of UNDEFINED speech acts. More examples for each type of speech act were provided, and the model was tasked with reviewing its own classification to find errors and identify more speech acts of specific types.

The final results are shown in the right column of Table I. Although the quantitative distribution of speech act types improved, there remained an asymmetry in the prevalence of representatives, paralinguistic units, and undefined forms, while the share of regulatives, directives, valuatives, suppositions, and commissives dropped sharply.

The work shows that ChatGPT is a promising tool for the automatic annotation of such complex linguistic units as speech acts. It performs well when annotating a small number of utterances, but errors arise when processing big data, which can be corrected with instructive prompts.

It was also observed that when tasked with classifying utterances, ChatGPT often forgets the context and must be explained the same things multiple times, though this doesn't always help. When using both a textual description (definition) of a speech act and a list of example utterances illustrating it, the model tends to rely on the latter, becoming more formal and failing to recognize related forms that were not included in the list. This creates the impression that the model switches to simpler formal text matching and forgets its analytical capabilities.

Future work on testing the model may proceed in the following directions: 1) determining whether the quality of speech act recognition depends on the sample size and what volume of text data can be considered the maximum allowable for one prompt, 2) more detailed study of the annotation results, analyzing the utterances themselves, and identifying weaknesses that prevent the model from making the correct interpretation, 3) optimizing prompts to achieve better results.

B. GigaCHAT

The work was conducted using the GigaChat model, version 3.5 with 13 billion parameters.

Step 1. The model understands the given task, supplementing its output with additional theoretical information from external sources.

Step 3. The model's output is not grounded in the given training examples. The class descriptions are based not on the provided examples but on external information.

Экспрессивные-эмотивные речевые акты выражают эмоции и чувства говорящего. Они могут быть использованы для выражения радости, удивления, разочарования и других эмоций. Примеры таких речевых актов включают восклицания, междометия и эмоционально окрашенные высказывания.

Рогативные речевые акты используются для обозначения действия или состояния, которое относится к другому лицу или объекту. Они часто начинаются с вопросительных слов, таких как "кто", "что", "где" и т.д., и направлены на получение информации от собеседника.

Вердиктивы-суппозитивы представляют собой утверждения, которые содержат предположение или допущение со стороны говорящего. Они могут использоваться для выражения мнения или суждения о каком-либо предмете или ситуации.

Step 4a. The model's analytical capabilities are constrained by its inability to identify multiple speech acts within a single utterance, often resulting in misclassifications.

1. Экспрессивный-эмотивный речевой акт: «*П мне / взять ?» — выражает просьбу.
2. Рогативный речевой акт: «а ? какой Vogue$? с ментолом //» — направлен на получение информации от собеседника.
3. Комиссивный речевой акт: «это вместе ? да ? не / вместе / вместе //» — устанавливает обязательства между участниками разговора.
4. Директивный речевой акт: «вместе ? потом рассчитаемся //» — направлен на то, чтобы заставить другого человека выполнить какое-то действие.
5. Паралингвистический речевой акт: «*Н //» — невербальное средство коммуникации, выражающее отношение говорящего к ситуации.
6. Вердиктивный-суппозитивный речевой акт: «что ещё ? всё //» — содержит предположение или допущение со стороны говорящего.

Step 4b. The version of the model available at the time of the research does not provide the ability to upload .xls files. To address this issue, three completed dialogues of different sizes were randomly selected from the general database.

During the analysis of the new data, the model continued to make errors both in identifying speech acts within a phrase and in determining their types, despite attempts at correction via prompts. Currently the model demonstrates an inability to handle such linguistic tasks, as its output results do not align with theoretical data and are random.

The problem of "forgetting" the context does not seem significant, as throughout the entire dialogue, the GigaChat model adhered to the given instructions, although it made multiple errors.

A promising direction for improving the model's performance may be to expand the training sample by adding a larger number of dialogues with human-annotated speech acts. This will allow the model to more effectively learn to solve the given task.

C. YandexGPT

The latest version of the model, YandexGPT 3, was used for the research.

Step 1. The model was tested with additional instructions for understanding the transcription notation and without it. The results showed no differences between the two cases. The model identifies the boundaries of speech acts situationally, without relying on the instruction.

Step 3. The model successfully generates descriptions of all classes based on the provided examples.

1. ЭКСПРЕССИВЫ-ЭМОТИВЫ
Этот класс включает в себя выражения эмоций и оценки ситуации. В данном случае, это фразы вроде «всё сделано / и всё так неплохо //», которые выражают удовлетворение результатом работы.
2. ВЕРДИКТИВЫ-СУППОЗИТИВЫ
Данный класс включает высказывания, содержащие оценку или суждение о чём-либо. В этом случае фраза «а-а / да это которую мы переделывали //» может быть интерпретирована как оценка качества работы или результата.
4. КОМИССИВЫ
Комиссивы — это обещания или обязательства. В этом примере фраза «потом рассчитаемся //» может рассматриваться как обещание оплаты или выполнения каких-либо обязательств.
6. ДИРЕКТИВЫ
Директивы — это указания или просьбы. В этом контексте фраза «хорошо? позвоните мне вечерком //» является просьбой позвонить в определённое время.

Step 4a. The model performed well on classifying a small number of utterances.

1. Мне сказали / что у меня какой-то хронический гайморит / и это может быть от этого // — неопределяемый фрагмент (фраза без контекста).
2. Может быть, от зуба / может быть, от зуба // — рогативы (запрос информации о причине проблемы).
3. А / даже так? — вердиктивы-суппозитивы (оценка вероятности предположения).
4. Угу // — экспрессив-этикетный (подтверждение или согласие).
5. А-а! — экспрессивы-эмотивы (выражение удивления).
6. *П — паралингвистика (использование невербальных средств коммуникации для передачи информации).*
7. А как вот её сделать? Это какие будут последствия? — рогативы (уточнение процедуры и последствий).
8. Резекция / да никаких последствий / единственное, что конечно, оно потом поболит после резекции / там дня два-три / может даже ра... припухнуть // — репрезентативы (представление фактической информации или описание ситуации).

However, due to the token limit in the prompt, the utterances were presented in small blocks, and the model failed to treat them as part of a continuous speech situation. As a result, it did not consider the context of previous messages, negatively affecting the quality of the responses.

Step 4b. YandexGPT does not support the upload of external files, so this step was skipped. All speech data was presented solely in the dialogue window.

The model performs well in generating class descriptions based on initial examples and in classifying a limited number of utterances. However, as the amount of text to be analyzed increases, the model becomes less accurate: it skips some speech acts, fails to classify them, and limits the number of recognized categories, focusing only on the most frequent ones instead of the full classification.

Another limitation of the model is its small working memory. While the dialogue context is retained, and the model can perform repetitive tasks without prompt duplication, it forgets the original data as the dialogue history grows, altering the classification based on new speech data, generating new classes, and removing less frequent ones.

Overall, YandexGPT3 is most successful at generating text and identifying the main idea, which allows the model to determine the illocutionary force of a speech act quite effectively. However, this focus prevents the model from classifying information according to predefined instructions, as it repeatedly generates and adapts classes to fit the specific linguistic material provided for analysis. These observations align with the statistics presented on the developers' website, indicating that information classification is the task with the lowest accuracy (59% compared to YandexGPT2).

As a result of this research, YandexGPT can be a useful tool for handling small-scale linguistic tasks, but the model is not suitable for working with large datasets and performs poorly in classification tasks.

To improve the model's performance and explore its potential further, the following areas can be considered:

1) How prompt optimization affects the model's performance,

2) The impact of refining initial classes on the quality of speech act annotation,

3) How the size of the training set (examples) for each class influences the model's understanding of speech act types.

D. Gemini

In this work, the Gemini family model, version 1.5 Flash, was used.

Step 1. The model demonstrates a high level of adherence to instructions, consistently fulfilling its expected role as a linguistic research assistant.

Step 3. Without explicit guidance, the model autonomously structures its output of class descriptions based on the provided examples, facilitating human understanding.

Эмотивные и этикетные экспрессивы

- **Экспрессивы-эмотивы:** Выражают субъективные эмоции говорящего по отношению к обсуждаемой теме или ситуации. Они могут быть как положительными (радость, удовлетворение), так и отрицательными (разочарование, недовольство).
- **Экспрессивы-этикетные:** Служат для поддержания социальных отношений и выражения вежливости. Они используются в приветствиях, прощаниях, благодарностях и других подобных ситуациях.

Вердиктивные и регулятивные

- **Вердикты-суппозитивы:** Высказывания, выражающие предположения, гипотезы или оценки о ситуации или предмете обсуждения. Они часто содержат модальные глаголы или частицы, указывающие на степень уверенности говорящего.
- **Вердикты-валюативные:** Выражают оценку качества, ценности или значимости чего-либо. Они могут быть как положительными, так и отрицательными.
- **Регулятивы коммуникативные:** Служат для организации и регулирования процесса коммуникации. Они используются для уточнения информации, подтверждения понимания, переключения темы и т.д.

Step 4a. The model accurately classifies short dialogues and provides additional explanatory notes. Gemini, without requiring additional prompts, structures the output in a tabular format that can be exported as an .xls file.

Реплика	Речевой акт	Подробное описание	Примечания
еще пожалуйста сигареты	Директив (просьба)	Прямая просьба о предоставлении сигарет.	Классический пример директива.
Vogue$	Репрезентатив (утверждение)	Упоминание конкретной марки сигарет как факта.	Подтверждение информации.
а ? какой Vogue$?	Рогатив (вопрос)	Вопрос, уточняющий марку сигарет.	Типичный вопрос для уточнения информации.
с ментолом	Репрезентатив (атрибуция)	Указание на свойство (атрибут) сигарет.	Описание характеристики предмета.
*П мне / взять ?	Рогатив (предложение)	Вопрос с предложением о действии.	Сочетание вопроса и предложения.
возьми	Директив (приказ)	Прямое указание на выполнение действия.	Более категоричная форма директива.
это вместе ?	Рогатив (вопрос)	Вопрос о суммировании стоимости.	Вопрос, связанный с финансовыми расчетами.
да ? не / вместе / вместе	Рогатив (вопрос), Репрезентативы (утверждение, отрицание)	Комбинация вопроса и утверждений, выражающих сомнение.	Сложный речевой акт, требующий контекста.
вместе ?	Рогатив (вопрос)	Повтор вопроса о суммировании стоимости.	Уточняющий вопрос.
потом рассчитаемся	Комиссив (обещание)	Выражение намерения совершить действие в будущем.	Обещание выполнить действие.
*Н	Неопределенный	Невозможно определить речевой акт без дополнительного контекста.	Требуется более глубокий анализ.
что еще ?	Рогатив (вопрос)	Вопрос о наличии дополнительных покупок.	Открытый вопрос, предполагающий продолжение диалога.
всё	Репрезентатив (утверждение)	Утверждение об отсутствии дополнительных покупок.	Закрывает предыдущий вопрос.
*П вам что(?) ?	Рогатив (вопрос)	Вежливый вопрос о желании собеседника.	Выражает заинтересованность в собеседнике.

Step 4b. The current version at the time of research does not support table uploads for analysis. Due to this limitation, three previously randomly selected dialogues were used for analysis. During the process, the model continued to demonstrate accuracy in classifying categories. However, despite having no explicit limitations on request volume, the model seemingly "economizes effort" and, while processing dialogues exceeding 20 lines, only works with the first 5-10 of them. Potential solutions involve prompt refinement. For instance, prompts specifying the analysis of "absolutely all" utterances yield the desired result. Additionally, working with one dialogue in smaller parts (less than 20 lines) with more specific prompts proves effective.

Requesting partial analysis without manual division was not efficient. While attempting to request an analysis "starting from the nth utterance" the model uses the context to generate hypothetical continuations rather than analyzing the rest of the actual dialogue.

Further work with Gemini 1.5 Flash may include 1) working with larger datasets and detailed output analysis to identify potential model weaknesses, and 2) further prompt specification to enhance efficiency.

C. Mistral

In the final stage of the practical part of the study was used the Mistral Large 2 model.

Step 1. The language model was provided with instructions to classify utterances into one of the speech act categories. The model was also introduced to specific symbols that were used for annotating the linguistic material. Mistral Large 2 readily adapted to the role of an assistant in linguistic research.

Step 3. The model successfully describes the categories, assigning each distinctive feature. It provides examples from contexts and speech situations that were used during training. The definitions provided by the model in the classification process appear to be accurate.

КОМИССИВЫ:

- Эти речевые акты включают обещания и предложения. Они выражают намерение говорящего выполнить определенные действия в будущем. Примеры: "потом рассчитаемся", "ну-ка", "давайте глянем".

ЭКСПРЕССИВЫ-ЭТИКЕТНЫЕ:

- Эти речевые акты связаны с вежливостью и этикетом. Они используются для выражения уважения, благодарности или других социальных норм. Примеры: "# очень (...) приятно", "здравствуйте".

ДИРЕКТИВЫ:

- Эти речевые акты включают просьбы и приказы. Они направлены на то, чтобы побудить собеседника выполнить определенные действия. Примеры: "хорошо? позвоните мне вечерком", "да", "свою забирайте".

РЕГУЛЯТИВЫ КОММУНИКАТИВНЫЕ:

- Эти речевые акты регулируют ход коммуникации. Они используются для управления разговором, уточнения информации или подтверждения понимания. Примеры: "@ ну вот", "я вам не сделаю не шестнадцать", "а восемнадцать сделаю", "хорошо?", "цветные матовые", "хорошо".

Step 4a. The model makes errors by not identifying more than one speech act within a single utterance. In challenging cases, it most often classifies the utterance as a 'representative', which is not always accurate.

1. **это вы сами мне / сдавали //**
   - **РЕПРЕЗЕНТАТИВЫ (ИНФ)**: Эта реплика передает информацию. Говорящий сообщает о том, что кто-то сдавал что-то.
2. **перед тем / как уехать в Голландию //**
   - **РЕПРЕЗЕНТАТИВЫ (ИНФ)**: Эта реплика передает информацию. Говорящий сообщает о времени перед уездом в Голландию.
3. **ну / пожалуйста / это ваше дело / У но теперь вы это / должны всё это / проанализировать //*
   - **ДИРЕКТИВЫ (ДИР)**: Эта реплика побуждает адресата к действию. Говорящий просит проанализировать что-то.
4. **дальше / да / следующая / смотрите / шкала вторая / спонтанная агрессивность //**
   - **РЕПРЕЗЕНТАТИВЫ (ИНФ)**: Эта реплика передает информацию. Говорящий сообщает о следующей шкале.

The model performs most successfully with the 'rogatives' category, paralinguistics, and others that point to extra-linguistic contexts. This is because utterances belonging to these categories often contain specific markers (e.g., '?', etc.).

Upon repeated requests to pay closer attention to the categories, the language model continues to perform in the same manner, classifying all unclear cases as 'representatives.' This suggests that when classifying speech acts, the model primarily focuses on the formal aspects of the utterance provided for analysis, without delving into semantic and syntactic relations or the grammar of the constructions, thereby overlooking the illocutionary force of the speech act.

Step 4b. The version available at the time of the study did not offer the option to upload tables for analysis. Due to this limitation, the same three randomly selected dialogues were used for analysis.

Thus, the Mistral Large 2 language model can be effectively used for annotating material that contains clear formal markers of a particular class, as more complex categories—determined by the sum of the semantic meanings of each individual component in a speech act—will not be accurately classified.

When attempting to break down an overly long speech act identified by the model into smaller speech acts, Mistral Large 2 responds by classifying both as 'representatives,' which is not always accurate.

The errors made by the model in classification tasks are likely due to its design and predominant use in English, unlike Russian-language models such as GigaChat and YandexGPT.

Further testing of the model can proceed in the following directions: 1) determining whether the quality of speech act recognition truly depends on the choice of material in a specific language, 2) identifying ways to better instruct the model, pinpointing weaknesses that hinder accurate analysis, 3) optimizing model queries to achieve more accurate results.

V. DISCUSSION

The study explored the comparative performance of five Large Language Models (LLMs) — ChatGPT, GigaCHAT, YandexGPT, Mistral, and Gemini — across four key criteria: capturing linguistic labels, refining outputs based on user feedback, handling large context windows, and processing data in Russian. Table II shows the summary of the results.

TABLE II. THE COMPARISON OF LLMS' CAPACITIES

Criterion	Chat GPT	Giga CHAT	Yandex GPT	Mistral	Gemini
CAPTU-RING LABELS	Requires refining	Struggles with linguistics	Better at short examples	Sticks to formal charac-teristics	Better at short examples
REFINING OUTPUTS	Efficient	Efficient	Efficient in short dialogues	Limited in some tasks	Efficient in short dialogues
CONTEXT WINDOWS	Better on shorter contexts	Relies on external sources	Limited	Large	Needs data chunking
PROCESS-ING RUSSIAN	High capacity	Trained in Russian, excels in Russian	Trained in Russian, excels in Russian	Limited	High capacity

A. Capturing Linguistic Labels

ChatGPT demonstrated moderate success in capturing linguistic labels, though a substantial portion of speech acts were initially labeled as UNDEFINED. By providing additional labeled examples, the model's performance improved significantly. GigaCHAT, however, struggled in this area, as it is more oriented towards solving common tasks rather than specialized linguistic analysis. In contrast, YandexGPT displayed strong performance in short examples with a high generalization capacity. Mistral, on the other hand, primarily focused on formal category representations, such as punctuation, and struggled with deeper contextual relationships. Gemini excelled in classifying short dialogues, even providing explanatory notes and structuring the output in a tabular format.

B. Refining Outputs Based on User Feedback

All models showed some level of refinement capabilities. ChatGPT benefited significantly from user feedback, resulting in improved labeling quality. GigaCHAT also successfully refined outputs based on feedback, while YandexGPT displayed high refinement capacity in short contexts but struggled with long ones due to context forgetting. Mistral, although capable of refinement in general tasks, showed limited success in linguistic categorization. Gemini demonstrated effective task refinement, particularly in dialogues, but only when the context was short; longer dialogues still posed challenges.

C. Handling Large Context Windows

The models showed varying abilities to handle large context windows. ChatGPT performed better with smaller datasets but tended to forget context during long refinement sequences. GigaCHAT mitigated context forgetting by using retrieval-augmented generation, which enhances answer quality but relies on external information rather than grounding itself in the training examples. YandexGPT's context window was limited, and it did not support external file uploads, restricting its use in large context scenarios. Mistral, while having a large context window, was hampered by its lower Russian-language capabilities, making conclusions on this aspect difficult. Gemini struggled with dialogue history memory, necessitating data chunking for high-quality analysis.

D. Processing Data in Russian

In terms of Russian-language processing, ChatGPT, GigaCHAT, and YandexGPT stood out. ChatGPT and GigaCHAT were both excellent in processing Russian, with GigaCHAT being particularly strong due to fine-tuning for this language. YandexGPT, similarly fine-tuned for Russian, excelled in determining the illocutionary force of speech acts. Mistral lagged in this area, showing lower capacity in Russian-language tasks. Gemini performed well, with high capacity in Russian and the ability to set a system role as a linguistic research assistant.

VII. Conclusion

The study describes an experiment on the automated speech acts annotation using five LLMs commonly used for processing Russian texts – ChatGPT, GigaCHAT, YandexGPT, Gemini, and Mistral. A comparative analysis of the automatic annotation results was conducted, highlighting the strengths and weaknesses of each model. The findings show that employing LLMs for corpus annotation is a promising method. While the automatically generated annotations are not without flaws and require expert revision, using large language models overall appears to be an effective tool for processing linguistic corpora.

This comparative analysis of these LLMs highlights their strengths and limitations across four key criteria. ChatGPT and Gemini demonstrated strong overall performance, particularly in processing Russian and refining outputs based on user feedback. GigaCHAT excelled in leveraging retrieval-augmented generation to minimize context forgetting, although it struggled with linguistic analysis. YandexGPT showed high capacity in short-context tasks and Russian-language processing but was limited by a smaller context window. Mistral, while capable in general tasks, was less effective in linguistic categorization and Russian-language tasks.

Overall, the study indicates that in the context of the automated linguistic corpora annotation task, ChatGPT and Gemini stand out for their adaptability in such complex tasks as speech acts categorization. However, when language-specific information is required, it is recommended to use language-specific models, such as GigaCHAT and YandexGPT for Russian. Future research could explore enhancing these models' capabilities in handling larger context windows and refining outputs over long sequences.

Acknowledgment

This article is an output of a research project “Text as Big Data: Methods and Models for Working with Large Textual Data” implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

References

[1] A. Asinovsky, N. Bogdanova, M. Rusakova, S. Stepanova, A. Ryko, and T. Sherstinova, “The ORD Speech Corpus of Russian Everyday Communication ‘One Speaker's Day’: Creation Principles and Annotation,” *Lecture Notes in Computer Science, Text, Speech and Dialogue*, vol. 5729, 2009, pp. 250-257.

[2] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, O. Ermolova, E. Baeva, G. Martynenko, A. Ryko, “Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech,” *Lecture Notes in Artificial Intelligence, LNAI*, vol. 9811. *SPECOM 2016,* Springer, Switzerland, 2016, pp. 659–666.

[3] D. Jurafsky, “Pragmatics and Computational Linguistics,” in *The Handbook of Pragmatics*, L. Horn and G. Ward, Eds., Oxford: Blackwell, 2006, pp. 578-604.

[4] M. Weisser, “Speech Act Annotation,” in *Corpus Pragmatics: a Handbook*, K. Aijmer and C. Rühlemann, Eds., Cambridge: CUP, 2014, pp. 84-111.

[5] T. Sherstinova, “Speech Acts Annotation of Everyday Conversations in the ORD corpus of Spoken Russian,” *Lecture Notes in Artificial*

Intelligence, SPECOM 2016, LNAI, vol. 9811. Springer, Switzerland, 2016, pp. 627–635.

[6] J.L. Austin, *How to Do Things with Words.* Oxford: Oxford University Press, 1962.

[7] J.R. Searle, “A Classification of Illocutionary Acts,” *Lang. in Society*, vol. 5, no. 1, 1976, pp. 1-23.

[8] A. Wierzbicka, “Akty mowy,” in *Semiotyka i Struktura Tekstu*, M.R. Mayenowej, Ed., Wrocław, 1973, pp. 201-219.

[9] M.M. Bakhtin, *Speech Genres and Other Late Essays*, C. Emerson and M. Holquist, Eds., V.W. McGee, Trans., Austin, TX: Univ. of Texas Press, 1986.

[10] Yu.D. Apresyan, “Performativy v grammatike i slovare,” *Izvestiya Akademii nauk SSSR. Seriya literatury i yazyka*, vol. 45, no. 3, 1986, pp. 208-223.

[11] T. Sherstinova, “Pragmaticheskoe annotirovanie kommunicativnykh edinic v korpuse ORD: mikroepisody i rechevye akty [Approaches to Pragmatic Annotation in the ORD Corpus: Microepisodes and Speech Acts],” in *Proc. Int. Conf. “Corpus Linguistics-2015,”* 2015, pp. 436-446.

[12] M. Weisser, “SPAACy: A Semi-Automated Tool for Annotating Dialogue Acts,” *International Journal of Corpus Linguistics*, vol. 8, no. 1, 2003, pp. 63-74.

[13] G. Leech and M. Weisser, “Generic Speech Act Annotation for Task-Oriented Dialogues,” in *Proc. Corpus Linguistics 2003 Conf.*, D. Archer, P. Rayson, A. Wilson, and T. McEnery, Eds., UCREL Tech. Papers, vol. 16, Lancaster Univ., 2003, pp. 441-446.

[14] J. Allen and M. Core, “Draft of DAMSL: Dialog Act Markup in Several Layers,” Web: https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/, 1997.

[15] S. Blum-Kulka and E. Olshtain, “Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP),” *Applied Linguistics*, vol. 5, no. 3, 1984, pp. 196-213.

[16] W. Stiles, *Describing Talk: A Taxonomy of Verbal Response Modes*, Newbury Park, CA: Sage, 1992.

[17] I.N. Borisova, *Russkiy razgovornyy dialog. Structura i dinamika [Russian Spoken Dialogue. Structure and Dynamics]*, Moscow: LIBROKOM, 2009.

[18] T. Sherstinova, “The Structure of Everyday Dialogue as the Sequence of Speech Acts,” in *Proc. Int. Conf. on Computational Linguistics and Intellectual Technologies, Dialogue 2018*, vol. 17, Moscow, Russian Federation, 2018, pp. 637-651.

[19] T. Sherstinova, I. Petrova, O. Mineeva, and M. Fedosova, “Empirical Studies of Everyday Professional and Domestic Communication for the Development of Voice Assistants in Russian,” in *Proc. 32nd Conf. Open Innovations Association (FRUCT), Tampere, Finland*, 2022, pp. 262-269.

[20] T. Sherstinova, S. Stepanova, A. Ryko, “Annotation system in the Russian language sound corpus ‘One speech day’”, *Proceedings of the XXXVIII international philological conference. Section: “Formal methods of Russian speech analysis”*, March 2009. SPbSU: SPb. Pp. 66–75.

[21] T. Sherstinova, “Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus,” in *SPECOM 2015, LNAI,* vol. 9319, A. Ronzhin et al., Eds., Springer, Switzerland, 2015, pp. 268-276.

[22] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is All You Need,” in *Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS'17)*, 2017, pp. 6000-6010.

[23] Achiam J. et al. “Gpt-4 technical report”, arXiv preprint arXiv:2303.08774, 2023, pp. 1-100.

[24] GigaChat, Web: https://giga.chat/.

[25] Yandex GPT 3, Web: https://ya.ru/ai/gpt-3.

[26] A.Q. Jiang et al. “Mistral 7B”, arXiv preprint arXiv:2310.06825, 2023, pp. 1-9.

[27] G. Team et al. “Gemini: a family of highly capable multimodal models”, arXiv preprint arXiv:2312.11805, 2023, pp. 1-90.

[28] R. Bommasani et al. “On the opportunities and risks of foundation models”, arXiv preprint arXiv:2108.07258, 2021, pp. 1-214.

[29] G. Marvin, et al., “Prompt Engineering in Large Language Models,” in *Proc. Int. Conf. on Data Intelligence and Cognitive Informatics*, Springer Nature Singapore, 2023, pp. 387-402.

[30] J. Schmidhuber, Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook. Technische Universität München, 1987.

[31] T.B. Brown et al. “Language models are few-shot learners”, arXiv preprint arXiv:2005.14165, 2020, pp. 1-72.

[32] P. Chen et al. “Iterative translation refinement with large language models”, arXiv preprint arXiv:2306.03856, 2023, pp. 1-10.