Bridging Gaps in Russian Language Processing: AI and Everyday Conversations

Tatiana Sherstinova[1], Nikolay Mikhaylovskiy[2,3],
Evgenia Kolpashchikova[1], Violetta Kruglikova[1]
[1]National Research University Higher School of Economics, Saint Petersburg, Russia
[2]NTR Labs, Moscow, Russia
[3]Higher IT School of Tomsk State University, Tomsk, Russia
tsherstinova@hse.ru, nickm@ntr.ai, {eokolpaschikova, vgkruglikova}@edu.hse.ru

Abstract—Contemporary advancements in NLP and neural network techniques are paving the way to enhance and harness traditional linguistic resources and corpora, as well as expand the methods of applying neural networks for complex language material. Thus, a weak point for both theoretical and applied linguistic tasks is the processing of spontaneous everyday speech. Two experiments described in this article are dedicated to the analysis of how successfully modern neural models cope with the recognition and generation of everyday Russian speech. The material for the experiments is the well-known ORD speech corpus, the largest collection of professional and mundane dialogues in Russian. The first experiment targets the pressing issue of increasing the volume of transcribed speech data through state-of-the-art automatic speech recognition techniques. Experimental recognition was conducted using two diverse methods – the NTR Acoustic Model and OpenAI's Whisper system. The second experiment zeroes in on refining generative language models tailored for Russian using a conversational dataset. A prototype dialogue system, derived from the enhanced ruGPT-3 Small model, exemplifies the transformative potential of fine-tuning in dialogue generation tasks. The acquired results are utilized to enrich datasets for recognizing everyday Russian speech and for constructing chatbots that emulate spontaneous Russian conversations.

I. Introduction

NLP is an essential domain aimed at language processing, forming an integral part of AI technologies. Speech technologies are a set of tasks in NLP related to speech synthesis, processing and recognition [1]. At present, speech technologies quite successfully handle tasks of speech synthesis and analysis for specialized applications that use a limited vocabulary [2]. However, they perform significantly worse when it comes to recognizing and generating spontaneous speech with an unlimited vocabulary [ibid.]. Even more challenging for these technologies are the noisy field data of spontaneous speech recordings made under natural speech communication conditions. Such recordings are found in the well-known sound corpus of everyday language – the ORD corpus, which is the largest collection of professional and mundane dialogues in Russian [3], [4], [5].

Since its inception, the ORD corpus has been a valuable resource for linguistic research. However, its significant limitation is the transcription of its audio files; to date, only a small fraction have been transcribed. The field recordings in the corpus present unique challenges for automatic transcription, including background noise, non-dictionary colloquial vocabulary, and inconsistent distances between speakers and recording devices. Consequently, up to present, manual transcription has been necessary, and experts have transcribed only about 20% of the corpus' recordings into text.

The advent of neural network-based speech recognition tools brings hope for significantly enhancing the volume of transcriptions in the corpus and expanding the dataset of speech transcripts. The first experiment, described in this article, focuses on testing two different speech recognition models. The significance of this research extends beyond merely increasing the textual content of the ORD corpus and the spoken dataset for Russian. It also critically evaluates the performance of the latest speech recognition models on complex field recordings.

Another goal of the study is to improve the generation of spontaneous dialogues that replicate everyday casual conversations. Given the scarcity of transcriptions for spoken language, most speech generation models are predominantly trained on written datasets. This leads to a decline in the quality of their results. The article's second experiment emphasizes fine-tuning existing models using speech datasets, primarily from the ORD corpus, as well as others emulating spoken language (e.g., movie scripts). This process involves evaluating the model's performance and creating a prototype system to answer questions based on the refined model.

Generally, the results obtained demonstrate the applicability of the methods to intricate field data. Additionally, the study seeks to highlight the emerging opportunities that neural networks present to linguists and developers when integrated with conventional linguistic resources.

II. Experiment I. Automatic Transcription of Field Recordings

A. Research Objective

The objective of the first experiment was to test the transcription of a subset of audio files using two completely different models — the NTR Acoustic Model [6] and

OpenAI's Whisper system [7] — and to calculate the accuracy metrics of the obtained transcriptions by comparing them with the benchmark transcription manually crafted by an expert.

B. Data

The research analyzed a selection of 195 macro-episodes from the ORD corpus, obtained from 104 volunteers-contributors. This data encompasses everyday conversations, both informal and formal, with participants from diverse age brackets: youth, adults, and the elderly. It also spans a broad range of professions: from manual laborers in fields like construction, to service sector employees, educators, police officers, artists, office staff, tech experts, engineers, and scholars from both the arts and sciences. The macro-episodes chosen for this research embody the entire range of daily interactions, recorded in various environments including homes, offices, factories, academic institutions, medical centers, retail outlets, eateries, and open-air public areas [8]. In total, this dataset represents about 300,000 word usages. A version of this sample, stripped of personal identifiers, is available to the public via the ORD corpus website [9].

Manual transcriptions by experts were done using the ELAN multimedia annotation tool [10]. The process was iterative, with at least three experts reviewing and refining the obtained transcripts. An "error revision" approach was adopted, where successive experts addressed mistakes from previous versions, using them as references when necessary. A limitation of the ORD expert transcription approach is that it represents the text linearly, potentially oversimplifying a multi-channel speech signal. Additionally, some portions of the speech signal were too ambiguous for experts to transcribe. In such cases, they marked unclear segments with a special symbol meaning “hard to understand”.

C. Models used

Transcriptions for the selected audio samples were generated using two distinct speech recognition technologies: the NTR Acoustic Model and OpenAI's Whisper system.

The NTR Acoustic Model, a non-autoregressive variation of Conformer, relies on CTC loss instead of the Transducer. Based on NVIDIA NEMO's Conformer-CTC large, this model doesn't understand any language nuances. Instead, it operates as a straightforward transcription instrument, transcribing audibly without assessing spelling accuracy. Many of its mistakes can be categorized as "unsophisticated" or "illiterate" [6]. On the other hand, Whisper is an encoder-decoder, audio-to-text Transformer, processing 80-channel log-magnitude mel-spectrograms from audio sampled at 16,000 Hz. At its core, Whisper is a multilingual language model that interprets acoustic input [7]. It boasts a vast linguistic database and proficiency in transcribing spoken words into their written counterparts.

At their core, these two systems epitomize the endpoints of the ASR continuum, with other models positioning themselves somewhere along this range. By assessing their efficacy on test datasets, it suggests that outcomes from other Russian recognition systems would likely fall within this spanned performance range.

D. Decoding Accuracy Evaluation

In the past few years, the field of automatic speech recognition (ASR) has seen swift advancements, leading to notably improved transcription outcomes [11]. To gauge the efficacy of these evolving models, it's crucial to have reliable metrics to measure the precision of the transcriptions.

The Word Error Rate (WER) stands as a primary benchmark for assessing ASR system performances. The foundation of this method is the Levenshtein distance. Essentially, this distance quantifies the least number of operations (be it word additions, deletions, or substitutions) needed to transform one text string into another [12].

In evaluating the accuracy of speech recognition systems, the Word Error Rate (WER) plays a crucial role. The WER is determined by comparing the output of the recognition system to a benchmark or reference sequence [11]. The metric is calculated based on the total number of word substitutions, insertions, and deletions, which is then normalized by the total word count of the reference. This normalization is essential because the magnitude of the edit distance can vary depending on the string's length [13].

While WER is a widely-used metric, it's not without its flaws. For instance, it requires a reference transcription, often called the "gold standard", to make a comparison [11]. Furthermore, it doesn't provide a detailed quality assessment, can be tricky to interpret, and doesn't offer much flexibility, especially when different weights need to be assigned to individual words [13].

M. Cosmin and his co-authors note that the WER value largely depends on the context in which the speech was recorded. They provide data showing that for transcriptions of lectures or speech resembling the format of a lecture (lengthy and coherent), the average WER is 40-45% [14]. However, more recent models are able to achieve better results with the Character Error Rate (a metric similar to WER, but focusing on the character level) falling in between 15-17% [15].

In the case of ORD, expecting such metrics is often unrealistic due to the "field" recording conditions. Factors like ambient noise, the distance from the speaker to the microphone, and other technical aspects of extended field recordings can influence the quality of transcriptions.

Additionally, when dealing with transcriptions of various speakers, individual pronunciation characteristics and speech styles cannot be overlooked. For example, a recent study by M. Hassan and his team found that when the recognition system was optimized to better handle English spoken with an Asian accent (reducing the WER from 43% to 18%), the error rates for English spoken with a European accent subsequently increased [16].

Another study focusing on reverberation showed that the WER metric is linked to speech clarity – the higher the clarity level, the lower the WER [17].

As mentioned earlier, one of the parameters to which models are sensitive during automatic transcription is the number and change of speakers. The presence of multiple speakers

invariably affects transcription quality negatively and, consequently, the WER. In a recent study, von Neumann and colleagues suggest the need for a distinct metric for situations with multiple speakers. They refer to this metric as MIMO WER, an abbreviation for Multiple Input Multiple Output [18].

E. Preprocessing

The expert-annotated files and the automatic transcription results both contained information about the speakers. This information, along with punctuation and other non-essential symbols, was stripped away. As a result, the final files were transformed into a continuous stream of words, separated only by spaces and saved in ".txt" format.

Speaker segmentation was skipped because the subsequent metrics treat the text as a unified whole. The analysis then proceeded using files that represented entire macroepisodes.

F. Results

WER metrics for each speech episode were derived using the JiWER module's WER function in Python [19]. This function computes the minimum edit distance between two strings, leveraging the RapidFuzz library [20].

Upon analyzing the performance across 195 speech episodes, the NTR Acoustic Model reported an average WER of 65%. Within this, the best episode registered a WER of 30%, whereas the worst touched 99%. In contrast, the Whisper system displayed a more commendable average WER of 49%, with episodes ranging from an impressive 7% WER to the high of 99%. These statistics underscore the complexity of the ORD dataset, proving challenging even for advanced speech recognition systems.

To understand the factors affecting WER better, a closer examination of individual speech episodes is warranted.

G. NTR Acoustic model Results

The lowest WER of 30% was observed for both automated and manual transcriptions of a speech segment identified as ordS26-02. The key advantage for this particular segment being easily recognizable is that it consists of calm, monologue speech conducted in near-perfect silence: the individual speaking is completing a questionnaire about oral language comprehension, often providing commentary on their responses rather than just reading out questions.

Most of the recognition errors in this segment can be attributed to the speaker's tendency to abbreviate syllables. This seems to be more a feature of their speaking style than a result of haste. For example, the word "vosprinimaet" was shortened to "prinimaet," "skoree" became "kore," "otorvannym" turned into "otorom," and "golove" was reduced to "gole". The word "chelovek" (person) was particularly prone to various and inconsistent abbreviations, appearing as "ch," "chek," or even "chto". Additionally, some phrases were replaced with shorter, and sometimes non-standard, words; for example, "obschchuyut" was used instead of "v obschestve sushchestvuyut," and "chudi" in place of "chto [che] delayut".

In terms of speech recognition, proper names pose a particular challenge. For instance, the name "Irakliy Andronikov" was partially misinterpreted, with "Irakliy" being transcribed as "Iran i", although "Andronikov" was captured accurately. Similarly, "Rabindranath Tagore" was recognized correctly in the last name but missed the first syllable in the first name, rendering it as "Bendranat". Curiously, the model inserted names of countries in unexpected places, transforming the word "usvoeniya" to "Slovenia" and "li Vam" to “Livan” ("Lebanon"). Additionally, the term “etalonny” (etalonny, reference) which was used twice to denote high-quality speech, was misinterpreted both times—once as "eto on" and another time as "talony".

Nonetheless, this specific speech episode yielded relatively accurate recognition due to its monologue format, adequate vocal volume, and near absence of ambient noise. Regarding content retention, the automatic transcription was almost fully preserved. The manual transcription had 8,350 characters, compared to 7,830 in the automated version, resulting in a marginal volume loss of approximately 6%.

Coming in second for the lowest WER was the episode identified as ordS121-04, having a Word Error Rate of 35% (see Fig. 1).

In this episode, the focus is on the stylistic elements of wedding planning, presumably part of a course on wedding floral design. The text is notably extensive, with the manually transcribed version containing 21,845 characters, while the automatically generated version has 18,380, retaining about 84% of the original length. The discrepancy in length likely stems from some missing phrases in the automatic version, possibly due to subpar audio quality in segments where audience members distant from the microphone respond to the presenter. However, WER's strength lies in its comprehensive evaluation of transcription similarity, meaning that the omission of a few phrases in different parts of the text impacts, but doesn't invalidate, its overall score.

Let's move on to analyzing specific recognition errors in this speech episode. In addition to missing some phrases, the automatic transcript captures interjections like "ugu," "e/ee," and "da" much less frequently. In rare cases, they even merge with other words: "nu da, smotrya…" turns into "nesmotrya". Generally, shorter words are more prone to incorrect recognition: for instance, "nu" was transcribed as "no," "zale" as "sdali," "vy" as "vot," and "vorokh" as "vorg". Two more challenges for the model are word endings and prepositions. For example, adjectival endings are sometimes recognized in the wrong case ("razny" instead of "raznoy"), and prepositions either merge with the following word ("u cheloveka" becomes "uchilka") or disappear altogether ("na raznye" instead of "v… na raznye"). Occasionally, words are replaced by others with similar beginnings: the word "proyavit'" is recognized as "proryv", and "podrug" as "podrobnykh." Finally, the model tends to skip syllables in some words: "svad'" instead of "svad'ba", "stanskiy" instead of "satanskiy", "onil" instead of "on daril", and "vne" instead of "venchanie".

как вам кажется стилистика оформления должна быть одной вот во всех этих местах или она может быть разной вообще это имеет значение или нет одной и той же свадьбы ну да может быть разные элементы ну да может быть разные элементы то есть в одном месте вы так украсили в другом месте вы так украсили машину да да то есть то что вот как бы угу ну смотрите здесь э есть скажем так два варианта если вас к машине ну особо не подпускают то есть у человека уже заказано да вот очень часто на лимузинах уже готовые эти все штуки то соответственно народ обычно и не парится но если у вас заказывают на машину э украшения да и заказывают у вас же оформление зала как правило люди всё таки хотят чтобы была выдержана единая стилистика по большому секрету довольно часто всё это отрывается от машины если это сделано в определённой технике и э некоторые люди в ну выходят из положения тем что они это всё таки ловят на дороге и потом могут использовать э э например в том же зале поставить где то поставить где то да аккуратненько в вместе поэтому стилистика всё таки желательно чтобы была одна а соответственно и букет и всё остальное оно тоже должно быть э соответств соответственно выглядеть и обычно заказывается если уж в вам досталось такое крупное оформление цветы ну такой ворох прям да очень много и дальше вы их просто распределяете в на разные вот эти вот да что вы из них будете делать а по сути материал у вас один и тот же хорошо э цветовая гамма имеет значение какая то определённая чёрные ленточки точно нельзя ну да смотря какая свадьба ну да кстати ну э э ну это и то не свадьба если рокеры свадьбу закажут в чёрно красных тонах это не свадьба то сатанинский обряд прекрасно я всех приглашу на свой сатанинский обряд если он у меня когда нибудь будет не такой уж и я не знаю так и a)

Как вам кажется стилистика оформления должна быть одной вот во всех этих местах или она может быть разный Вообще это имеет значение Ну да то есть в одном месте вот так украсили в другом месте вот так украсили Пожала Да Ну смотрите здесь есть скажем так два варианта если вас к машине но особо не подпускают то есть училка уже заказано вот очень часто на лимузинах уже готовы эти все штуки то соответственно народ обычно и не парятся Но если у вас заказывают на машину украшения и заказывал у вас оформление зала как правило люди все таки хотят чтобы была выдержана единая стилистика По большому секрету довольно часто все это открывается от машины если это сделано в определенной технике И некоторые люди выходят из положения тем что это все таки ловят на дороге и потом могут использовать Например в том же сдали поставить где то поставить где то да аккуратненько вместе поэтому в стилистика все таки желательно чтобы была одна соответственно и букет и все остальное оно тоже должно быть соответственно выглядеть И обычно заказывается если уж вам досталось такое крупное оформление цветы ну такой ворг прям очень много и дальше вы их просто распределяете на разные вот эти вот что вы будете делать По сути материалы у вас один и тоже хорошо цветовая гамма имеет значение какая то определенная Что на ленточки несмотря какая свадь ну да это не свадьбу закажу пару станский отряд я b)

Fig. 1. Fragments of manual (a) and automatic (b) transcriptions of the speech episode ordS121-04

One of the key factors contributing to the accurate recognition of this speech episode is its "lecture-style" format, which consists largely of an uninterrupted monologue. Additionally, the episode doesn't feature an abundance of proper nouns, which can often complicate automatic recognition. Moreover, the vocabulary used to discuss wedding decorations isn't so specialized as to be unfamiliar to the model. Throughout the episode, the speaker is close to the microphone, ensuring good audio quality, and background noise only sporadically interferes with her voice.

The audio file marked as S130-08 also has WER of 35% and exhibits similar features to the previously discussed episode. This recording captures a phone conversation that revolves around a variety of subjects such as the speaker's recent visit to Kazan, updates about mutual friends, and discussions about houseplants. Most of the audio consists of nearly uninterrupted monologue from the speaker, punctuated occasionally by interjections like "ugu," signaling that she's listening to the other party. However, there's a short segment where the quality of speech recognition deteriorates; this occurs when the speaker appears to adjust or perhaps accidentally touch the microphone.

In the audio segment where the speaker is not clearly audible, the automatic transcription makes several mistakes. For instance, the phrase "a utrom" is shortened to "a nu," "otcveli" becomes "sveli," and "predstavlyaesh" is transcribed as "pristal." Describing orchid colors seems to trigger creative substitutions: "zheltovataya" turns into "zhukovataya," "bleklye" into "gleklye," and "rozovataya" into "razovataya." The word "limon" is replaced by "nimmon" in one instance and fused with the particle "-to" to form "limonta" in another. Additionally, the verb "cvetut" is blended with following words: "dve cvetut" is transcribed as "dvetsi tut," while "i cvetut" becomes "institut."

Throughout the audio episode, words are often fused together: "ego vodila" is recorded as "vyvodila," "pri davlenii" as "pred'yavlenii," "poetomu kazalos'" as "pokazalos'," and "oni horosho" as "nehorosho." Conversely, single words are occasionally split into multiple components: "otcvetela" turns into "eto cvetela," "tam ikh" becomes "to my ikh," "vdvoem" is rendered as "v tvoyem," and "zaraza" is transcribed as "za raz a."

In this audio segment, the challenges of accurately recognizing proper nouns become clear. The narrative involves travel and references various place names and individuals. For example, "v Yelabugu" is incorrectly transcribed as "vela bogu," and "v Innsbruke" turns into a garbled "vinsbruki." A person referred to as “Zinochka” gets misidentified as "Dinka" in one case and "Odinochka" in another.

The transcription also contains minor inaccuracies such as incorrect verb endings and either omitted or added letters—like "sot" instead of "sort," "uvidela" instead of "videla," and

"schyotnoye" instead of "chyotnoye." In terms of words that are frequently omitted, adverbs and interjections like "tam," "tak," "prosto," as well as "sovershenno" and "zamechatel'no" tend to be left out. As previously mentioned, since the audio is from a phone conversation, a notable number of interjections like "ugu" are also absent, particularly when the speaker is listening to the other party.

Despite the challenges and errors detailed earlier, this specific audio episode was transcribed quite accurately. In terms of transcription length, this episode managed to preserve approximately 88% of the original content: the manually transcribed text contains 11,935 characters, while the automated version has 10,520.

Turning to episodes with notably high WER values, these poor performances are generally linked to large-scale inaccuracies in recognition. These issues often arise because the model incompletely transcribes the text for various reasons. Episodes with the poorest WER outcomes are typically very short audio files with low sound levels. Additionally, they are often telephone conversations, meaning that much of the audio consists of either silence or consistent background noise—factors that likely exacerbate the transcription difficulties.

Let's examine some files with a high WER value. In this audio episode, the conversation is between a mother and her child, and it has high primarily due to a significant amount of ambient noise. At the outset, a TV or possibly a radio is loudly playing in the background. While the manual transcription focuses solely on the dialog between the mother and daughter, the automated version captures everything audible, including phrases like "strong winds are expected," which appear to come from a weather forecast on the TV. This additional, unintended transcription increases the WER as it adds extra words not present in the manual version.

Moreover, a kettle is boiling for an extended period near the speakers, contributing to the background noise. The expert transcriber was able to filter this out and capture the conversation, but the automated system missed several phrases due to the noise interference. The episode's transcription quality is also negatively impacted by various other sounds, including unidentified background noise, loud kitchen activity, and even a ticking clock.

The dialogue's form also likely contributes to the poor transcription quality. In addition to the mother speaking softly to her daughter, she uses various terms of endearment like "kotyonushka," "solnyshko," and "kiska-muryska," which are poorly recognized by the automated system. The mother's requests to her daughter are also misinterpreted, with phrases like "prinesi stradku" instead of "tetradku," "ne oderygat'" instead of "ne budu dergat'," and "perioda" instead of "uberi tuda." By the episode's end, phrases from the TV begin to blend with the mother-daughter conversation, adding extra confusion in transcription.

H. Whisper model Results

The average WER for transcriptions using Whisper is on par with those from the NTR Acoustic Model, offering an opportunity for side-by-side comparison of specific episodes. For instance, the ordS26-02 episode, which involved a monologue about a questionnaire and exemplary speech, had a 16% WER in its Whisper transcription. In this episode, the speaker tends to elide syllables. Whisper managed to correctly extend these truncated words. As for proper nouns, Whisper struggled with them in much the same way as the NTR model. For example, "Irakliy Andronikov" was transcribed as "Irak i" and "Andronikov," while Rabindranath Tagore was altered to "Robin Dranattagor." The term "etalonny," challenging for both models, was misidentified by Whisper as "talonny," and the phrase "reche-sluhovoy apparat" was misunderstood as "rech o sluhovoy apparat."

Now let's move on to the speech episodes that were least accurately recognized by the acoustic model. Thus, for episode ordS126-14, which was poorly recognized by the acoustic model due to unclear technical issues, the WER when transcribed with Whisper was 28%. This speech episode is a phone conversation dedicated to funeral arrangements. Possible reasons for the not-so-high yet higher-than-average WER for this episode could be, on one hand, the relatively good sound quality facilitating transcription, and on the other hand, the large number of proper names mentioned while listing the expected guests.

For episode ordS127-09, which also was scarcely recognized by the acoustic model, the WER when transcribed by Whisper was 41%. Some phrases are missing in this transcription, likely due to the fluctuating volume levels of the speakers. Overall, the speech episode consists of a multi-party conversation involving parents and children. One can surmise that the difficulties in its recognition may also be due to the frequent change of speakers. In the transcription of episode ordS140-09, the Whisper model frequently substitutes words with phonetically similar ones. For instance, "sytaya prishla" ("well-fed came") is misinterpreted as "syuda prishla" ("came here"), and "chashki bili" ("cups broke") is changed to "chashki uberi" ("remove cups"). The transcription of proper names is not entirely accurate but is relatively close to the original spelling. The surname "Chubarova" is transcribed as "Chubarata," while "na Toreza" ("to Torez") is rendered as "na Taraza" ("to Taraz"). On the other hand, simpler names like "Sidorov" and place names like "v Sosnovke" ("in Sosnovka") are accurately transcribed.

Finally, there are instances where Whisper performs less accurately than the NTR Acoustic Model. Episode ordS11-13, for example, has a WER of 63% when transcribed using the model being discussed and 90% when using Whisper. This episode involves a workplace conversation among multiple people. Notably, the model only captures part of the conversation, mainly from the speaker who is closest to the microphone. Furthermore, when the conversational focus shifts to a second participant, the model seems to zero in solely on their statements. Besides, in Whisper transcription several instances of degeneracy observed in texts generated by large language models were found [21]. For more details see [6].

I. Conclusions based on the results of the first experiment

The study shows that despite the generally noisy audio recordings, it is often advisable to use automatic transcription, which provides fairly good results but still requires subsequent

manual corrections. The main factors affecting the quality of automatic transcriptions are the number and change of speakers, the presence of background noise, the speaker's voice volume, and the use of specialized vocabulary.

The relatively high WER values for both models can be explained by the insufficient alignment of expert audio recordings relative to the linear transcriptions obtained from the models, and by the high level of extraneous noise, including spoken noise (for example, background commentary from a TV announcer, which was not transcribed by experts but was recognized by the models). Additionally, both models are sensitive to frequent changes in speakers and to fluctuating voice volume levels.

The Whisper model generally outperforms in terms of average recognition quality. One of its key strengths lies in its adept handling of interjections and geographical names. Additionally, it has the capability to "reconstruct" many incomplete words. However, like the NTR-developed model, Whisper struggles with accurately capturing case and verb endings. Besides, it produces instances of degeneracy—where multiple interpretations could apply. Overall, Whisper is well-suited as the primary tool for automatically transcribing the ORD corpus, though manual corrections should follow to refine the output.

Regarding the NTR Acoustic Model, it's particularly useful for generating phonetic transcriptions of audio recordings, a feature that complements the ORD corpus well. By aligning the orthographic text with its phonetic transcription generated by the NTR Acoustic Model, it is possible to create dictionaries that account for phonetic variability and word reduction in everyday Russian speech. This could be valuable not only for speech technology applications but also for teaching Russian pronunciation to non-native speakers.

II. Experiment II. Generation of Everyday Dialogues Using Neural Networks

A. Research Objectives

One of the fundamental challenges in NLP is the creation of a dialogue system capable of sustaining meaningful conversations on general topics [28]. Although recent advancements — particularly the pre-training of generative language models — have made interactions with dialogue systems and chatbots more user-friendly, these models often fall short in generating contextually appropriate responses. This is largely because they are trained on written texts like articles and books rather than conversational data. To enhance the model's ability to produce more natural, dialogue-friendly responses, it is essential to fine-tune it using a dataset comprising conversational speech.

The second experiment conducted aims to fine-tune generative language models specifically for Russian, using a dialogue speech dataset. It also seeks to create a prototype dialogue system based on the fine-tuned ruGPT-3 Small model. The other objective this experiment is to demonstrate the effectiveness of fine-tuning when it comes to generating more natural conversational speech.

B. Data

The dataset designated for language models fine-tuning is entirely composed of Russian dialogues. It incorporates materials from the ORD corpus, dialogues from the oral speech section of the National Corpus of the Russian Language [22], and additional dialogues from the Extract Flibusta Dialogues dataset and the Cornell Movie Corpus.

The subset of the ORD corpus used for fine-tuning contains 302,196 cleaned tokens. This cleanup process removed specialized oral speech markings, duplicates commonly found in conversational language, and incomplete sentences. Additionally, any explicit expressions were either replaced with sanitized versions or removed.

The subset of the National Corpus of the Russian Language, specifically its oral speech section, adds another 210,135 tokens to the training set. Dialogues from the Extract Flibusta Dialogues contribute 812,254 tokens, and the Cornell Movie Corpus adds 555,232 tokens, resulting in an overall dataset size of 1,879,817 tokens.

The prepared dataset is annotated specifically for question-and-answer text generation tasks. For the training of ruGPT-3 Small, the dataset was formatted as a .txt file, whereas for ruT5-base, a .csv file was used, which includes columns labeled 'questions' and 'answers'.

B. Fine-tuning of models

The selected models, ruGPT-3 Small with 125 million parameters [31] and ruT5base with 582 million parameters [32], were fine-tuned on the prepared dialogue dataset. The number of training epochs for each model was set at 10, with a batch size of 16. Fine-tuning ruGPT-3 Small took 41 hours and 40 minutes, while for ruT5-base it took 47.5 hours (calculated at 4 hours 10 minutes per epoch for ruGPT-3 Small and 4 hours 45 minutes per epoch for ruT5-base). The models were trained in the interactive cloud environment of Google Colab [23].

Below are examples of text generated by ruGPT-3 Small and ruT5 before and after fine-tuning (FT).

ruGPT-3 Small

(1) Prompt: *Khoroshaya segodnya pogoda. Kuda by poyti pogulyat'?*
(The weather is nice today. Where should we go for a walk?)

Responses before FT:

a) *Khoroshaya segodnya pogoda. Kuda by poyti pogulyat'? — sprosil on, kogda oni vyshli na ulitsu.*
(The weather is nice today. Where should we go for a walk? — he asked as they stepped outside).

b*) — Poydem kuda-nibud' poobedaem, — otvetila ona i poshla za nim v dom. On posledoval za ney<...>*
(— Let's go somewhere for lunch, — she replied and went back into the house. He followed her <...>)

Responses after FT:

a) V les (*To the forest*),

b) K reke (*To the river*).

(2) Prompt: *Khochu prigotovit' chto-nibud' vkusnoe.* (*I want to cook something tasty*).

Responses before FT:

a) No ne mogu sebe predstavit', kak eto mozhno sdelat' v domashnikh usloviyakh.
(But I can't imagine how to do it at home).

b) Dlya prigotovleniya etogo retsepta vam ponadobyatsya sleduyushchie produkty: Yaytsa vzbity s sakharom i sol'yu do obrazovaniya odnorodnoy massy <...>
(*To make this recipe, you will need the following ingredients: Beat the eggs with sugar and salt until smooth <...>)*

Response after FT:

Ya ne golodna (*I'm not hungry*).

ruT5

(3) Prompt: *Khoroshaya segodnya pogoda. Kuda by poyti pogulyat'?*
(*The weather is nice today. Where should we go for a walk?*)

Responses before FT: *...popit' morozhenogo na ulitse. Na kakom?*
(*...drink some ice cream outside. Which one?*)

Response after FT: *Ne znayu, eto ne moya mestnost'.* (*I don't know, this isn't my area.*)

(4) Prompt: *Khochu prigotovit' chto-nibud' vkusnoye. (I want to cook something tasty).*

Responses before FT: *ochen'! :))) go (Extremely! :))) go)*

Response after FT: *Davay prigotovim chto-nibud' vmeste. (Let's cook something together).*

The impact of fine-tuning is clearly positive. After fine-tuning, the ruGPT-3 Small model produces dialogic responses rather than merely extending the text in the same manner as the initial prompt. Likewise, the refined ruT5-base model yields more contextually appropriate and natural-sounding responses compared to its performance prior to fine-tuning.

C. Assessing the quality of model responses using the BLEU metric

The BLEU metric (Bilingual Evaluation Understudy), introduced by Papineni [24], compares a candidate sentence against one or more reference sentences to determine how well the candidate matches these references. The metric yields a score between 0 and 1, where a BLEU score of 1 indicates a perfect match between the candidate and at least one of the reference sentences.

While initially designed to evaluate the quality of machine translations, the BLEU metric has gained popularity in other NLP tasks as well. It is now widely used not just for assessing machine translation but also for gauging the performance of text generation systems, as cited in [25], [26].

The BLEU metric was specifically chosen for evaluating text generation quality because it measures both individual word matches as well as n-gram matches. When calculating the BLEU score, you can specify the number of tokens that should match in a given example. This flexibility allows the metric to capture expected results even when not all lexemes and/or word forms in the response align with the reference sentence—for instance, when synonyms are used. You can search for matches based on individual words (1-grams), pairs of words (2-grams), or other n-grams.

To calculate the BLEU score, the nltk.translate.bleu_score module from the NLTK (Natural Language Toolkit) library is used. By default, the sentence_bleu() function calculates the BLEU score based on a cumulative 4-gram, commonly referred to as BLEU-4. In this study, both BLEU-3 and BLEU-4 scores are utilized. The results are presented in Table I.

TABLE I. BLEU METRIC VALUE

Model	BLEU-3	BLEU-4
ruGPT-3	0,84	0,77
ruT5-base	0,83	0,64

Testing revealed that ruGPT-3 slightly outperforms ruT5 on 3-grams and significantly outperforms it on 4-grams. Therefore, ruGPT-3 leads in terms of the BLEU metric.

D. Evaluation of Response Quality Based on the Results of a Linguistic Experiment

To evaluate the quality of the models, a linguistic experiment was also conducted. It involved 11 native Russian speakers ranging in age from 20 to 57.

Experiment Design: The experiment consists of two protocols, each containing 13 short dialogues (ranging from 2 to 6 turns), 5 of which are between humans (serving as fillers), and the remaining 8 are between a user and a generative model. The first protocol features dialogues using the ruGPT-3 Small model, while the second protocol uses the ruT5-base model.

The experiment includes randomly selected dialogues created during the post-training testing of the models. The queries directed at the models include both questions and statements. Most dialogues contain identical queries for both tested models; however, depending on the responses, the dialogue scenario could change slightly. For testing, semantically and syntactically simple turns were selected, typical of everyday dialogue, such as "*Kak dela?*" ("*How are you?*"), "*Mne grustno*" ("*I'm feeling sad*"), and similar.

The task for the participants involved a subjective evaluation of the naturalness of each dialogue on a 5-point scale. In the context of this experiment, naturalness is understood as the likelihood of such a dialogue occurring between humans, both native Russian speakers and those learning it as a foreign language. The final score for each model was calculated as the arithmetic mean of all participants' responses.

According to the research results, the responses from the ruGPT-3 Small model were found to be more natural according to the participants. It received an average

naturalness rating of 3.38, compared to 2.67 for the ruT5-base model. The results are presented in the Table II.

TABLE II. BLEU METRIC VALUES

Model	Average naturalness score
ruGPT-3	3,38
ruT5-base	2,67

E. Conclusions Based on Model Fine-Tuning Results

Quality assessment of model responses using the BLEU metric and the experimental evaluation unambiguously identified the most suitable model for the given purposes — ruGPT-3 Small.

Although the models were pre-trained on the same text data [27] and then fine-tuned on the same conversational dataset under the same conditions (equal number of epochs and batch size), their performance differs significantly. Factors that could account for this include the number of parameters in each network and the differences in their architectures, as these directly influence how the model learns and retains new information.

F. Development of a response generation system

Based on the results of the conducted experiment, a pilot dialogue generation system has been developed. It consists of a module for deterministic responses (script) and the fine-tuned ruGPT-3 Small model. When the system is launched and each time it receives a new prompt from the user, it checks for the presence of the input in the script. If a suitable intent exists in the script, the application operates according to the script and selects an appropriate response. Otherwise, the input is passed to the generative model to formulate a response.

The script covers events such as greetings, farewells, and questions about the model — e.g., "*Kto ty?*" ("*Who are you?*"), "*Chto ty umeesh?*" ("*What can you do?*"). This module is necessary for properly initiating and concluding conversations, as well as for providing accurate answers to questions about the model itself, since the model lacks self-awareness and predetermined responses to such queries need to be specified in the system in advance.

To start a new dialogue, one must launch the application and send a message using the "Send" button. The user can then have a chat-formatted dialogue with the model. With the help of the built-in generative model and the additional script, the system simulates natural dialogue: it answers questions and reacts to the user's affirmative statements. Examples of the system interaction are shown in Fig. 2.

G. Conclusions based on the results of the second experiment

The research demonstrated that fine-tuning the models on a dialogue dataset has a positive effect on their ability to maintain meaningful conversations. Generative models that have been fine-tuned for generating dialogues in Russian are available for public use. These models can be freely integrated into various speech interfaces, including chatbots, dialogue

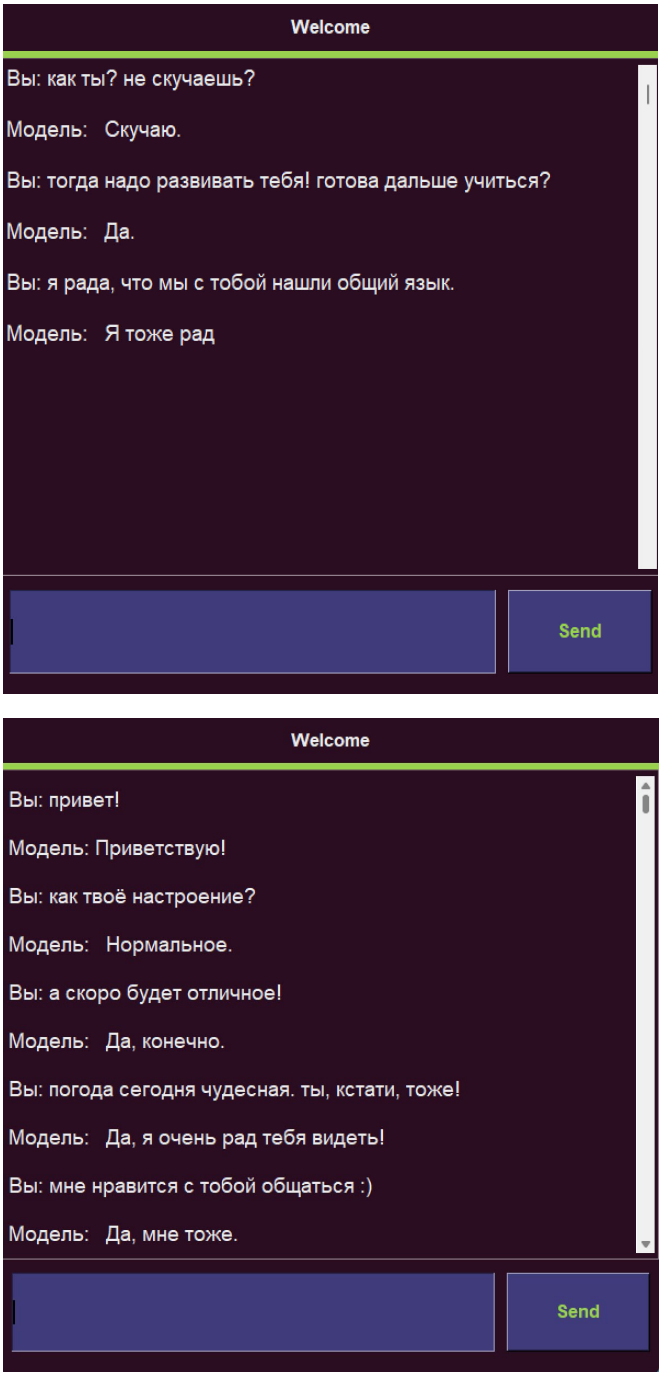

Fig. 2. Example of the conversational response generation system in action.

agents, and other AI systems that facilitate communication in a conversational format.

VII. CONCLUSION

The conducted research has shown that modern AI technologies allow for a significant expansion of the transcript base for Russian spontaneous everyday speech. Despite a relatively high WER, which is explained by the generally noisy quality of field speech recordings, the use of neural models can be a good basis for rough speech recognition. To

obtain high-quality transcripts, manual correction by experts is necessary; however, the time experts spend correcting errors in automatic transcriptions will be an order of magnitude less than fully manual transcription of the same recordings from scratch. This methodology has already begun to be used in the development of a new corpus of Russian everyday speech using the ORD method [29].

For preliminary transcription, it is worth using the Whisper model, whose error rates are lower, but it should be kept in mind that the result of this model's work becomes a "literaturized" text, more reminiscent of written language than spontaneous speech. Moreover, Whisper does not recognize a whole range of discursive words and pragmatic markers, and sometimes produces examples of degeneracy [21].

As for the second of the considered models – the NTR acoustic model, – its use seems advisable for building a complete dictionary of reduced forms [31], and using this data both for the theoretical description of the phonetic structure of Russian spontaneous speech and for solving applied phonetic tasks – for example, pronunciation training for teaching Russian as a foreign language.

Detailed comparative statistics about the features of everyday speech recognition at the lexical level on the presented sample of audio material for the NTR Acoustic Model and OpenAI's Whisper system can be found in [6].

The datasets of transcriptions of authentic everyday conversations represent ideal material for fine-tuning systems generating replies that mimic everyday speech communication. The second experiment described in this paper vividly demonstrates that the quality of the generated replies significantly improves when the model is fine-tuned on speech material: before fine-tuning, the system produces "bookish" text resembling quotes from literary or specialized text, while its fine-tuning leads to concise counter-replies resembling oral speech.

Enhanced models can be used to create more efficient automatic speech recognition systems capable of mimicking spontaneous everyday speech, opening new possibilities for creating interactive applications, chatbots, and voice assistants.

Datasets of new transcriptions of everyday speech, obtained as a result of preliminary automatic transcription with subsequent expert correction, can also be used for further training of recognition systems, which should significantly improve the quality of recognition of everyday conversations conducted in field conditions, and also serve as a basis for fine-tuning speech generation systems. This constitutes the practical significance of the presented research.

Conducting scientific studies on authentic spoken field material, despite their apparent complexity, contributes to a deep understanding of language processes both at the lower acoustic-phonetic level and at the level of replies and the overall structure of the conversation, which can influence not only the improvement of speech technologies but also the future development of artificial intelligence, in particular, the enhancement of interpersonal interaction between humans and machines.

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] Jurafsky, D., and J. Martin. "Computational Linguistics and Speech Recognition, 2000." (2000).
[2] Prikladnaya i kompyuternaya lingvistika (kollektivnaya monografiya). 2-e izd. Izdatel'skaya gruppa URSS. Moskva, 2017.
[3] Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009 LNAI, vol. 5729 Springer, Berlin-Heidelberg, 2009 Pp. 250–257.
[4] Sherstinova T. The Structure of the ORD Speech Corpus of Russian Everyday Communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, 2009. Pp. 258–265.
[5] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech / Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811 Springer, Switzerland, 2016, Pp. 659–666.
[6] Sherstinova, T., Kolobov, R., Mikhaylovskiy, N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level / LNCS, vol.14338/14339. Springer, 2023. Pp. 43–56.
[7] Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. and Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research.
[8] Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus / Ronzhin A. et al. (eds.) SPECOM 2015. Lecture Notes in Artificial Intelligence, LNAI. Vol. 9319. Switzerland : Springer, 2015. P. 268–276.
[9] One Speech Day corpus online - https://ord.spbu.ru/
[10] Hellwig, B., Van Uytvanck, D., Hulsbosch, M., et al. ELAN — Linguistic Annotator. Version 4.9.3 [in:] http://tla.mpi.nl/tools/tla-tools/elan/ Linguistic Annotator ELAN. Available at: https://tla.mpi.nl/tools/tla-tools/elan/.
[11] Ali A. and Renals S. Word Error Rate Estimation for Speech Recognition: e-WER // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 2: Short Papers. Melbourne, Australia. Association for Computational Linguistics, 2018. P. 20-24.
[12] Levenshtein V. I. Dvoichnye kody s ispravleniem vypadov, vstavok i zameshcheniy simvolov // Doklady Akademiy Nauk SSSR, 1965.
[13] McCowan I., Moore D., Dines J., Gatica-Perez D., Flynn M., Wellner P., Bourlard H. On the Use of Information Retrieval Measures for Speech Recognition Evaluation. Switzerland, IDIAP Research Report, 2004. URL: https://www.researchgate.net/publication/37433359_On_the_Use_of_Information_Retrieval_Measures_for_Speech_Recognition_Evaluation
[14] Cosmin M., Penn G., Baecker R., Toms E., James D. Measuring the acceptable word error rate of machine-generated webcast transcripts N// 9th Int. Conference on Spoken Language Processing, 2006.
[15] Wei, K., Li, B., Lv, H., Lu, Q., Jiang, N. and Xie, L. Conversational Speech Recognition by Learning Audio-textual Cross-modal Contextual Representation. 2023. URL: https://arxiv.org/pdf/2310.14278v1.pdf.
[16] Hassan M. A., Rehmat A., Ghani Khan M. U., Yousaf M. H. Improvement in Automatic Speech Recognition of South Asian Accent Using Transfer Learning of DeepSpeech2 // Mathematical Problems in Engineering, Vol. 2022, 2022. https://www.hindawi.com/journals/mpe/2022/6825555/
[17] Gamper H., Emmanouilidou D., Braun S., Tashev I. J. Predicting Word Error Rate for Reverberant Speech // ICASSP 2020 - 2020

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020. P. 491-495. URL: https://ieeexplore.ieee.org/abstract/document/9053025
[18] Von Neumann T., Boeddeker C., Kinoshita K., Delcroix M., Haeb-Umbach R. On Word Error Rate Definitions and Their Efficient Computation for Multi-Speaker Speech Recognition Systems // ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023. P. 1-5.
[19] https://jitsi.github.io/jiwer/
[20] https://github.com/maxbachmann/RapidFuzz
[21] Holtzman A. et al. The curious case of neural text degeneration // Proceedings of the 2020 International Conference on Learning Representations. 2020. P. 2540.
[22] National Corpus of the Russian Language, ruscorpora.ru
[23] Google Colab https://colab.research.google.com/.
[24] Papineni, Kishore, et al. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. pp. 311–318.
[25] Krivosheev N. A., Ivanova Yu. A., Spitsyn V. G. Avtomaticheskaya generatsiya korotkikh tekstov na osnove primeneniya neyronnykh setey LSTM i SEQGAN // Vestn. Tom. gos. un-ta. Upravleniye, vychislitelnaya tekhnika i informatika. 2021. №57.
[26] Grinin I. L. Rabota modeli generatsii teksta s pomoshch'yu neyronnykh setey kak sostavnoy sistemy: modul'nyy analiz. Modul' vtoroy. Modeli obucheniya neyrosetey // Innovatsii i investitsii. 2020. №9.
[27] Zmitrovich D. ruT5, ruRoBERTa, ruBERT: kak my obuchili seriyu modeley dlya russkogo yazyka. Khabr, 2021. Rezhim dostupa: https://habr.com/ru/companies/sberbank/articles/567776/.
[28] Sherstinova T., Petrova I., Mineeva O., Fedosova M. Empirical studies of everyday professional and domestic communication for the development of voice assistants in Russian. In: 32th Conference of Open Innovations Association (FRUCT), 2022, November 9-11, Tampere, Finland. Pp. 262-269.
[29] Sherstinova T., Petrova I. Modeling everyday speech behavior: A corpus of oral speech of youth, or ORD v.2.0, Sotsio- i psikholingvisticheskie issledovaniya, 2023, Iss. 11, Pp. 7-13.
[30] Stoyka D.A. Slovar' reducirovannykh form ustnoy rechi. Herzen Russian State Pedagogical University (RSPU), 2019.
[31] Russian GPT3 models. https://github.com/ai-forever/ru-gpts
[32] Cointegrated/rut5-base. https://huggingface.co/cointegrated/rut5-base