Empirical Studies of Everyday Professional, Domestic and Client-Service Communication for the Development of Voice Assistants in Russian

Tatiana Sherstinova, Irina Petrova, Olga Mineeva, Maria Fedosova
National Research University Higher School of Economics, Saint Petersburg
Saint Petersburg, Russia
tsherstinova@hse.ru, {iapetrova_2; oamineeva; mvfedosova}@edu.hse.ru

***Abstract*—Voice assistants are gradually becoming an increasingly common feature of our everyday life. However, the naturalness of communication provided by them usually leaves much to be desired. It may be caused by the fact that many chat-bots are trained on artificially created linguistic data sets and on fictional dialogues modeled by linguists on the basis of common phrasebooks or communication textbooks. As a result, the necessary pragmatic result can be achieved, but the feeling of "unnatural" communication of a voice assistant remains, which often reveals itself by the use of archaic phrases or remarks that are not quite suitable for the situation. This state of affairs seems to be improved by referring to real speech data — namely, to a representative volume of sound recordings of real speech communication. The paper discusses some approaches to the analysis of speech data from the sound corpus "One Day of Speech", which is the most representative resource of Russian everyday spoken communication. The pragmatic structure of professional and everyday conversations is considered, as well as linguistic content of standard modules, such as Greeting and Farewell. As a practical recommendation, we can suggest increasing the variability of answers not due to the lexical diversity of phrases, but due to a more diverse intonation implementation for the most typical replicas in spoken Russian.**

I. Introduction

More than 3 billion people turn to voice assistants daily. The number of Google assistant [1] users by 2020 has exceeded 500 million people [2], about 8 million users talk daily to Alice [3] — voice assistant from the Russian company Yandex. Apple's Siri [4] and the hundreds [5] of lesser-known customer-service communication, sociology, and education apps that hit the market every month have come a long way over the past decade. At the same time, projecting bots with human traits and the ability to communicate at the level of a real native speaker is becoming not just following the trend, but a necessity for developers [6]. That could be the main way to gain a competitive advantage on the market and to force users to resort to voice assistants.

In the case of using a voice assistant in the B2B and B2C segments, the unique language identity of the development and the most realistic simulation of live communication is the key to popularity in the market and the way to build customer trust and loyalty. When using bots in education, their ability to maintain a conversation on a variety of topics and at the same time to respond in a variety of ways is necessary to immerse the student in the realities of real communication.

Companies wishing to bring their voice assistant to the international market are faced with the task of teaching it to speak fluently in the appropriate language, adhering to the norms of modern communication. This happened with localization in Russia and the CIS countries, for example, with Siri voice system, for which Russian is not a native language. According to the observations of Russian-speaking, and not English-speaking users, for whose needs the listed assistants were originally created, the result of "relocating" of these bots to the Russian market seems not to be ideal. Unlike Alice, whose native language is Russian, foreign interfaces often do not maintain many communicative "plots", and cannot always successfully recognize Russian speech and its nuances outside of a clearly defined framework. Even such a popular assistant with "foreign citizenship" as Siri, which is used by most IOS users, is often able to use no more than 4-5 answers even to a typical question.

However, the naturalness of communication provided by state-of-the-art voice assistants usually leaves much to be desired. It may be caused by the fact that many chat-bots are trained on artificially created linguistic data sets and on fictional dialogues modeled by linguists on the basis of common phrasebooks or communication textbooks. As a result, the necessary pragmatic result can be achieved, but the feeling of "unnatural" communication of a voice assistant remains, which often reveals itself by the use of archaic phrases or remarks that are not quite suitable for the situation. Below are some examples:

- *Zdravstvuyte, kak pozhivayete?* [Hello, how do you do?]
- *Rad vstreche!* [Glad to meet you!]
- *Moyo pochteniye!* [My respects / My admiration!]
- *Budu zhdat' vas s neterpeniyem!* [I look forward to seeing you!]

The situation seems to be improved by referring to real speech data — namely, to a representative volume of sound recordings of real speech communication. We believe that one of the largest corpus of the sound Russian-language speech known as "One Day of Speech" [7; 8; 9] can help expand the base and become material for the further development and

more successful localization of the already popular voice assistants of foreign companies and those voice assistants that would be introduced to the market. Due to the variety of the collected data and its existence in audio format, this corpus is seen as one of the richest sources for the developments in the field of speech technologies. The paper discusses some approaches to the analysis of speech data from this well-known speech resource.

II. The ORD corpus of Russian everyday speech and its pragmatic annotation

A. Principles of ORD creation and collection of audio data

Russian linguists began to collect data obtained on the basis of the 24-hour recording technique for the "One Day of Speech" corpus in 2007 [7]. Each of informants-volunteers of the project was given a voice recorder. Participants were asked to turn on the device in the morning and, if possible, keep the audio recording on during the entire day from waking up to sleeping. Similar method of collecting authentic speech data is traditionally used in Japanese linguistic research [10; 11], it has also been used when collecting audio data for the Demographic Subcorpus of the British National Corps [12].

As a result, the recorded speech reflects the whole variety of communicative situations in which a person took part during one day being in different social roles. Over the years of work, linguists have collected a representative amount of speech data and managed to process more than 1450 hours of speech recordings received from 128 informants and more than 1000 of their communicants, representing different social groups of the modern Russian city. The volunteers represent a well-balanced sample, in which the number of men and women, people of different ages, different social origins and professions, whose speech was recorded, is balanced, and the resulting sound recordings demonstrate the diversity of speech behavior of the inhabitants of the modern Russian metropolis [8; 28]. The spontaneity of participants' speech during the experiment was not limited by any rules: there were no special requirements for the place or time of conversation, or the style of communication, or the quality of the recording. Due to this, the ORD corpus contains the most natural examples of spontaneous colloquial Russian speech, which allows it to be actively used to solve many theoretical, practical, and applied linguistic tasks.

One of these tasks is to study the features and structure of everyday speech communication in different real-life conditions. To support these studies, special pragmatic annotation of the corpus is carried out, which involves its division onto *macroepisodes*, *microepisodes*, and *speech acts*.

B. General principles of ORD pragmatic annotation

An initial and mandatory stage of audio processing when creating the corpus is an expert listening to the entire array of audio recordings, the removal of long pauses that do not contain speech, and the subsequent manual segmentation of audio files into *macroepisodes*, which are the main units of data storage and processing. The average duration of macroepisodes is about 15-20 minutes. Most of them are relatively homogeneous speech episodes, united by location, participants, and the main communicative task of communication. Each macroepisode gets normalized verbal description in three following aspects: (1) *Where does the situation take place?* (2) *What are participants doing?* (3) *Who is (are) the main interlocutor(s)*. The annotation scheme for macroepisodes is described in [13].

In the process of further annotation each marcoepisode is divided into microepisodes [14]. Microepisode is a relatively finished communicative fragment, homogeneous from the point of view of its main pragmatic task or conversation topic. It is microepisodes that are the closest prototype of a person's targeted conversation with a virtual assistant, since his/her goal in most cases is to solve a specific pragmatic task, which is most often the only one. If there are more such tasks than one, their consistent solution is assumed with the implementation of two or more microepisodes.

Microepisodes consist of separate utterances, each of which, in turn, has a certain generalized pragmatic meaning, expressed in referring them to certain speech act categories. There are a large number of different theoretical schemes of speech acts, starting from the classical works of J. Austin [15] and J. Searle [16], as well as practical schemes for marking speech material used in different languages [17-21].

To annotate sound recordings of Russian speech, it was decided to use speech act classification scheme proposed by I. Borisova [22], which was significantly revised. The main types of speech acts in the ORD corpus are defined as follows [23]:

- **Representative speech acts** (INF) are speech acts, the main purpose of which is the exchange of information between the participants of the dialogue.
- **Directive speech acts** (DIR) encourage the addressee to action (or inaction) or express an attempt to influence his worldview, emotions, and attitudes.
- **Commissive speech acts** (COM) are associated with the adoption of certain obligations by the speaker.
- **Emotive speech acts** (EMO) are used to express and convey feelings and emotions.
- **Etiquette speech acts** (ETI) are standardized forms that regulate communication in etiquette and ritualized situations.
- **Evaluative speech acts** (VAL) are used to express evaluative opinion or opinion-judgment.
- **Suppositive speech acts** (SUP) are used to express the opinion or assumption of the speaker.
- **Regulative speech acts** (REG) are phatic speech acts associated with the "organizational" aspects of the interaction, used to structure and conduct a dialogue.

A specific feature of any real speech communication is the fact that some of the utterances remain incomplete or "cut off" in mid-sentence. In this case, spoken fragments do not always make it possible to reconstruct the original intention of the speaker, that is, what type of speech act was planned to be reproduced. In the corpus, such speech segments are marked with a code (FRA), which means an undefined fragment — that

is, an incomplete speech act, by which it is impossible to determine its illocutionary force.

Most of speech acts are subdivided into more detailed categories, with a total of more than 150. Moreover, the list of these categories is not closed. The need to involve a detailed list of categories is justified by the fact that it is necessary, in particular, to separate "questions" from "answers" in the general category of *representatives* or "request" from "order" among *directives*. The most frequent forms of these subtypes are presented in [14].

III. The Structure of Everyday Conversations

A. Exploring the structure of everyday conversations

The study of Russian everyday dialogue structure was carried out on the material of 73 microepisodes of everyday speech communication from the ORD corpus [24]. The task was to find out what types of speech acts initiate and complete a dialogue most often, as well as to identify the most typical sequences of speech acts in the dialogue structure.

To achieve this goal, speech of 30 people (6 informants and 24 their interlocutors) was analyzed in the amount of 2230 speech acts related to both professional and everyday conversations. To calculate the most frequent sequences of speech acts, the technique of n-gram analysis was used.

The study showed that almost 40% of all statements belong to the category of representatives, the main task of which is to exchange information, 12.5% of speech acts are regulative forms that set and support the course of the dialogue, 11.4% of all statements are evaluatives expressing evaluative opinions or judgments. Directives (6.7%), etiquette forms (4.2%) and other categories of RA are much less common, finally, 9.38% of all statements in the sample are of mixed types (see Table I).

TABLE I. Distribution of the main types of speech acts

#	Types of Speech Act	Abbreviation	Abs. number	Percentage %
1	Representatives	INF	884	39,62
2	Regulative forms	REG	279	12,51
3	Evaluatives	VAL	254	11,39
4	Directives	DIR	151	6,77
5	Etiquette forms	ETI	93	4,17
6	Paralinguistic forms	PAR	83	3,72
7	Emotives	EMO	79	3,54
8	Commissives	KOM	62	2,78
9	Suppositives	SUP	58	2,60
10	Unfinished fragment	FRA	57	2,55
11	Undefined fragment	NER	18	0,81
12	Mixed types	INF/EMO, INF/REG, INF/VAL, DIR/ETI, etc.	249	11,16

Among the representatives, the most frequent categories turned out to be questions (28.65%) and explicatives (explanations, 17.35%), the most common regulatory forms are various kinds of boundary markers (beginning, segment, end – "tak" [so], "vot" [here, well], etc.), speech support like "aga", "ugu", "da" [yeah, yeah, yes], etc., and re-question. Among evaluatives in pilot sample, the most common turned out to be expressions of consent or approval (37.84%), while the objection occurred almost 4 times less often (9.96%). The most frequent directives are offer (25.83%) and request (19.87%). Among etiquette forms, vocatives (a third of all implementations) and greetings (28%) stand out, whereas the most typical emotives are positive emotions (21%) and surprise (14%). Commissives and suppositories are used rather infrequently in speech communication and their subtypes are few: for commissives, this is an agreement to fulfill a request (30%), a statement of intent (30%), and a promise (12%); for suppositives, an assumption (65%) and an expression of personal opinions (23%) are the most frequent [24].

It was also shown what types of speech acts start or end a conversation most often. Thus, in 16% of cases, the typical beginning of a conversation is a question, in 17% of cases, a microdialogue is initiated by a marker of the beginning of a new topic (for example, by "tak" [so], which is often being used in large macroepisodes). Dialogues also regularly begin with a greeting (13%), a message (11%), or a vocative (10%).

As for the utterances that complete dialogues, there were almost no regularly repeated ones observed at the level of speech act subtypes. The most frequent ending of the dialogue turned out to be agreement, which ends the dialogue in 10% of cases. Besides, statements and regulatory forms (for example, "vot" [here]) are relatively frequent (4% of cases each). Thus, it turned out that the beginning of a dialogue seems to be much easier to be formally predicted and modelled than its end [ibid].

B. Studying the structure of client-service communication

Since the use of voice assistants is mostly focused on solving client-service tasks, the next stage of research was dedicated to studying this type of everyday macroepisodes. The following main types of client-service communication can be distinguished in ORD data: 1) purchases, 2) obtaining information, 3) medical services, 4) briefing/instruction, and 5) provision of other services (repair, printing, ordering, etc.).

For a test sample of 10% of ORD client-service conversations, the preliminary distribution of speech acts by type of client-service communication is as follows: 1) purchases — 47.77%, 2) obtaining information — 31.26%, 3) medical services — 9.28%, 4) briefing — 6.86%, and 5) provision of other services — 4.82%.

A pilot sample of speech microepisodes of this type has been manually annotated on the level of speech acts, Table II reveals the comparative distribution of speech act main categories for different types of conversations.

TABLE II. Distribution of speech acts in client service communication in general and for particular types of conversation (Shop — purchases, Inst. — instruction, Med. — medical consultations, Inform. — information services)

#	Speech Act Types	Totally	Type of Client Service Communication			
			Shop.	Instr.	Med.	Inform.
1	Representatives	68,49	68,60	**75,00**	64,04	68,48
2	Regulatives	10,42	**7,93**	6,82	2,25	**14,79**
3	Etiquette forms	7,29	**10,04**	6,82	–	5,84
4	Directives	6,12	5,18	**11,36**	**17,98**	2,72
5	Evaluatives	2,60	3,05	–	4,49	1,56
6	Emotives	2,47	2,44	–	8,99	1,17
7	Suppositives	1,43	1,52	–	–	2,33
8	Commissives	1,17	1,22	–	2,25	0,78

The comparison of the most frequent subtypes of speech acts in everyday speech in general and in client-service communication is as follows (see Table III).

TABLE III. THE MOST FREQUENT SUBTYPES OF SPEECH ACTS WHEN COMPARING EVERYDAY PROFESSIONAL AND DOMESTIC CONVERSATIONS VS. CLIENT-SERVICE COMMUNICATION

#	Subtypes of speech acts	Common everyday speech, %	Client-service communication, %
1	question	11,26	**19,12**
2	answer	4,71	14,73
3	informing	5,34	**8,53**
4	notification		5,30
5	explication	6,77	4,65
6	statement		4,52
7	correction		3,75
8	gratitude		3,49
9	speech support	1,84	**3,49**

Table III shows that questions, informing and speech support are used more often in client-service communication than in other settings of everyday speech. From pragmatic point of view, this is quite understandable. As for the "answer" category, despite the evidently different percentage, the comparison is difficult here, since a slightly different methodology was used when marking up this data. In particular, when annotating client-service episodes, it was decided that any response after a question in client-service communication should be tagged as an "answer", while when marking everyday speech, different categories of answers could be possible to appear. However, due to the not very large size of the studied sample, the obtained quantitative data should be considered preliminary.

C. Studying the structure of buyer-seller communication

As it was shown in the previous section, the greatest amount of customer-service communication refers to the verbal interaction between buyer and seller in the context of purchasing. Therefore, this type of communication has become the subject of a separate investigation.

For this study, a subcorpus corresponding to the topic "shopping" was compiled. It consists of 57 episodes, each of them has a duration of 1 to 30 minutes. The total number of annotated speech acts is 4465 (2229 of them belong to clients and 2236 to personnel). There are 15 main categories of speech acts in the subcorpus, which in turn are detailed into 136 subcategories.

When analyzing the overall distribution of speech acts types it was observed that the majority of all utterances (almost 60%) are representatives. Regulative forms account for 17%. They are followed by evaluatives, directives, and etiquette expressives — approximately 5% for each type.

Five the most frequent speech acts categories in buyers-sellers communication correspond to the distribution speech acts obtained for ORD data earlier (according to [24]). The following data differ. Thus, in the subcorpus of buyer-seller communication, the percentage of commissives and is higher than it was obtained for everyday speech on average. According to ORD data, emotives and paralinguistic events are in these positions instead of commissives and suppositives respectively.

Such distribution of speech act types may be explained by specific communication aims. For successful interaction between customers and sellers, the verbal expression of one's intentions which is represented by commissives, is as essential as providing information. Emotionality, in contrast, is reduced since social roles of client and service employee do not involve personal relationships between people, which often entail freer expression of emotions as well as the use of paralinguistic phenomena (e.g., laughter). Social role of the customer or service employee is one of the factors that determines the nature of communication, therefore it seemed appropriate to have a separate look at the percentage of each type of speech acts for buyers and sellers (see Table IV).

Table IV presents the overall distribution of speech acts categories between these two roles. Representatives, or statements containing information are in the first place for both cases. They are followed by regulative forms which are needed to organize the conversation (e.g., "aga" [uh-huh], "ponjatno" [I see]). The third position in the clients' speech is occupied by etiquette forms ("zdravstvuyte" [hello], "spasibo" [thank you]), while directives are more typical for service employees ("podozhdite" [wait], "obratite vnimanie" [pay attention], "seychas, smotrite" [now look]). The next most frequent speech act category for both roles are evaluatives ("mezhdu prochim, tozhe ne ochen" [by the way, it's not very good either], "normalny kombez" [normal suit], "ochen', kstati, interesnyj" [very interesting, by the way]). Representatives and etiquette expressives are obligatory speech acts for each episode of buyer-seller communication in our sample.

TABLE IV. DISTRIBUTION OF SPEECH ACTS CATEGORIES IN BUYER-SELLER COMMUNICATION, %

#	Speech Act Types	Speaker's Role			
		Client	Rank	Service	Rank
1	Representatives	27,9	1	31,4	1
2	Regulatives	10,5	2	6,8	2
3	Evaluatives	2,6	4	3,0	4
4	Directives	2,0	5	3,1	3
5	Etiquette forms	3,1	3	1,7	5
6	Commissives	1,1	6	1,5	7
7	Suppositives	1,0	8	1,6	6
8	Emotives	1,1	7	0,4	8
9	Unfinished/ undefined fragments	0,4	9	0,1	10
10	Paralinguistic forms	0,1	10	0,2	9
	Totally	49,8	-	50,2	

The difference in speech acts frequency reflects the influence of speaker's role on his/her speech behavior. For example, the prevalence of directives in the speech of service employees reflects the specific tasks of salespersons to offer, recommend goods, and guide the customer. The role of the client is usually more passive and does not imply particular functions that influence communication. At the same time, in most cases it is important for customers to respect the accepted norms of communication and behavior. Thus, 3% of customers' speech are etiquette forms.

Buyer-seller communication relates to one of the most usual activities for any person. The results obtained demonstrate the existing conventional, and in certain cases institutionalized,

rules of communication in service sector. The findings illustrating the empirical distribution of speech acts in conversations between buyers and sellers are essential in order to maintain the naturalness when developing communication models of virtual assistants that correspond as closely as possible to the reality.

One of the most common examples of accepted rules of communication is its beginning and its end. According to our sample, in a customer-seller interaction, the client more often (in 60% of cases) greets first and then, without waiting for a response, asks a question. In this case, a salesperson directly responds to the request. However, if the seller initiates the communication, the buyer will nearly always greet him or her in return.

The end of the dialogue is commonly characterized by a customer's gratitude, which is followed by a response from the salesperson (22%, such as "pozhalujsta" [you're welcome]), or another client's question (8%). After seller's response goodbye comes (8%), which, in contrast to the greeting subtype, is more characteristic for the role of a service employee, then for the role of a client (52%).

IV. Etiquette Forms in Everyday Business and Professional Communication

A. Greetings and goodbyes etiquette forms

In the final part of the paper, we would like to focus our attention on a relatively special, but very important aspect of communication associated with the beginning and ending of conversations, namely greetings and farewells [25]. As shown in Section III, 13% of microepisodes within macroepisodes begin/end with a greeting/farewell. For independent mini-dialogues between sellers and buyers, this proportion is even higher.

During real communication, the opinion on the interlocutor is largely formed particularly by his first and last replicas. Similarly, in a conversation with a modern voice assistant, the user tends to evaluate its "level" by how humanly it starts a conversation, and can maintain and finish it. Because of this, we believe that it would be logical to start improving voice assistants precisely with the listed elements — greetings and farewells.

Talking about everyday communication and so called client-service conversations, it can be noted that due to the change of interlocutors in the communicative flow throughout a day, elements of professional, client-service, and domestic types of communication are repeatedly encountered. Of course, in the case of real everyday communication, the speakers' replicas are characterized by greater variability, while in the situation of client-service-oriented communication the speakers are forced to adhere to the rules of business etiquette. However, it seems that when teaching a general conversational system (such as Alice, Siri or Alexa) on this material, the latter can safely use both more formal and informal vocabulary and mix them in a free order in the intents system. General voice assistants are not as limited by formalities as the apps used as sales or service consultants. At the same time, the selected materials, if necessary, can also be used for teaching the latter [26]. To do this, the development team would only have to use the data related exclusively to the business type of communication.

When selecting the empirical material used for the development of voice assistants, it was decided to focus on the approach to everyday communications and conversations between individuals acting in a role of client and a service worker. As empirical observations show, examples of such communication become an invariable part of almost any day of each ORD informant. As a result, these elements of communication are evenly distributed over the wakefulness of people, regardless of their age, gender, and social status.

B. Greetings and goodbyes forms used by the popular Russian voice assistants

Testing of "Salut", "Siri" and "Alice" virtual assistants has shown that all of them react by using a greater variety of cues at the beginning of a conversation. At the same time, it should be noted that some answers of voice assistants differ from each other only in word order or by the use of synonyms. Here, a good example could be two greeting forms used by "Salut" voice assistant: "Spasibo, i vam vsego dobrogo" [Thank you, and all the best to you] and "I vam zdorov'ya, spasibo" [Good health to you, thank you].

1) The most frequent mistakes in smart assistant applications

When discussing mistakes made by voice assistants, it should be noted that the cases of inadequate response to a user's requests were found while testing Siri and Google Assistant. Thus, in response to "good morning", Siri does not use any greeting form at all. Instead, it informs that Apple's smart home platform Homekit, which is designed to user control of various internet-connected home devices, is unavailable at the moment. As for Google Assistant, it often does not take into account the time at which the conversation with the user takes place. This leads to a situation where, when talking in the morning, the system does not correct the user in response to "good evening", as other voice assistants usually do, but also uses the same greeting form "good evening". As for the "Salut" application, it was found that it is not able to recognize the greeting in the modern colloquial form of the word "Privet!" [Hello], which sounds like "Hayushki" [Hey-hey] or "Hay" [Hi]. These colloquial forms have been popular among young native Russian speakers in recent years. Both of them originate from informal English greetings [27].

2) When assistants tell jokes

The debate about whether artificial intelligence can tell jokes has been going on for a long time, and some researchers believe that artificial intelligence is not yet capable of doing so. The main reason for that is the lack of creativity as well as any kind of thinking process in AI inner structure. Although, it is worth saying that despite the limitations of artificial intelligence, voice assistants, whose creation is done by native speakers, are able to create a semblance of a joke. When creators include typical idioms and phrases from films and books to its database, some assistants sound human. In this case AI could mix senses and operate words the same way native speakers do. For example, "Salut" operates with "kiber-privet" [cyber-hello] emphasizing that the user still

communicates with artificial intelligence. In the morning, Alice is prone to use the colloquial phrase “Utro dobrym ne byvayet” [“There is nothing good about mornings”] popular among native Russian speakers. At the same time in response to “Goodbye”, this voice assistant often asks users “ne szhigat' mosty” [“not to burn all the bridges”]. Given the fact that in the mentality of Russian native speakers the expression “burn all the bridges” means “completely change everything / change one's life and break from the past”, we can say that artificial intelligence asks the user to use it again as soon as possible in such an unusual way.

3) Archaisms in the voice assistant vocabulary

The outdated database of active vocabulary seems to be one of the main voice assistant disadvantages. Old phrases and expressions that are currently unpopular among native speakers do not allow voice systems to be on the same page with users. Moreover, this very fact could be an issue when using voice assistants for teaching Russian. Such expressions as “privetstvuyu” [I welcome], “priyatno bylo pogovorit” [it was nice to talk], “moyo pochteniye” (my admiration), ”zdravstvuyte, kak pozhivayete?” [Hello, how do you do?], “budu zhdat' vas s neterpeniyem” [I look forward to seeing you] in modern Russian can be used in an extremely narrow usage. Their use in a wider context and everyday practice is rather limited. In most cases, native speakers use such words in an ironic way just to emphasize the comical nature of the situation.

Taking up the idea that modern voice assistants should be as human-like as possible, we suppose that the variety of using replicas is an obvious advantage for any voice system. It seems to be the best way to build a wide inner database for a voice assistant using examples of real modern speech without archaisms or any kind of old-fashioned vocabulary. Ideally, the voice system should be developed in accordance with the real language on a daily basis. That is why the process of including the most “trend” and modern phrases in the base of its intents seems to be useful for many applications. However, new words used by young people seems good mainly for assistants of general format, in chat bots intended for promotion of goods and services (for example, in the banking sector) it is worth to use more common vocabulary.

C. Description of empirical data from the ORD corpus

Studying speech recordings from the ORD corpus, we selected 115 audio files, among which there were 53 macroepisodes of client-service communication and 62 macroepisodes of everyday domestic communication. In this sample the numerous cases of polite and more informal forms of greetings and farewells were found.

Predictably, standard phrases used at the beginning and end of communication (such as greetings “Zdravstvuyte!” [Hello!], “Privet” [Hi], “Dobroye utro / Dobryy den' / Dobryy vecher” [Good morning / Good afternoon / Good evening]) are the most common (see Table V). These phrases are included in the vocabulary of many voice assistants.

As mentioned earlier, voice assistants need to imitate human speech, and achieve maximum naturalness and variety.

TABLE V. THE MOST WIDELY-USED GREETINGS AND FAREWELLS IN ORD CORPUS DATABASE

	Type of communication		Total number	%
	Client-service	Domestic		
Greetings				
Zdravstvuyte [Hello! / How d'ye do?]	34	11	45	75.6 / 24.4
Privet [Hey!]	16	18	34	47 / 53
Dobroye utro / Dobryy den' /Dobryy vecher [Good morning / Good afternoon / Good evening]	10	4	14	71.4 / 28.6
Farewells				
Do svidaniya [Goodbye]	26	6	32	22.6 / 5.22
Poka [Bye]	4	6	10	40 / 60

To achieve speech diversity most of them extend bot vocabulary with outdated words or phrases. However, in natural speech diversity is achieved by other methods. Thus, typical greetings and farewells are often accompanied by interjections and various particles [28]. For example, “Nu, privet” [Well, hello], “Aga, poka” [Yeah, bye], “Ladno, do svidaniya” [Okay, goodbye]. Further, some function words like “davay” [come on] can completely replace the standard form of farewell. You can often hear people say “Nu, do svidaniya” [Well, goodbye] or “Aga. Davay!” [Yeah, go ahead / Yep, see ya!]. Most voice assistants rarely, if ever, use such forms.

Thus, empirical data show that in real life, contrary to expectations, not very large set of utterances of the beginning and end of a conversation is used. The range of utterances of well-known voice assistants is more representative due to the use of archaic and specially invented phrases. Apparently, this is one of the factors preventing us from perceiving such communication as natural.

What significantly distinguishes natural communication from the imitated one is the variety of intonations used in real life and the adaptation of lexical and prosodic characteristics of the interlocutors' voice messages to each other (in particular, to the rate of speech) observed in many cases. It seems that it is in this direction that voice assistants should be developed, claiming the naturalness of the generated communication.

V. CONCLUSION

The paper describes an empirical study conducted on the base of real-life everyday conversations — professional, domestic, and client-service, — from the point of view of identifying their structural features that can be used to improve the quality of client-oriented voice assistants, in particular, to increase such a significant parameter of chat bots, as “naturalness” of the generated communication [29]. The results obtained should be taken as preliminary, it is desirable to continue the study with the involvement of a larger amount

of annotated data. However, taking into account the high labor intensity of manual speech tagging, a significant increase of empirical data would be rather difficult.

Besides, it seems appropriate to mention other difficulties that researcher encounter when studying samples of everyday speech recorded in real-life settings. They are, in particular, the following:

- When determining macroepisode types using the accepted scheme, there is often a deviation from the "ideal" client-service pattern. For example, the number of participants in the conversation may be more than two, their involving and role in dialogue may change (e.g., there may be a change of consultant or the involvement of several specialists at the same time), a heterogeneity in the structure of the conversation may also take place, when other lines of communication are superimposed on the solution of the main communicative task.
- In some cases of real everyday communication, in contrast to the simulated artificial scheme, the meaning of some statements may not be clear without a wider context.
- When studying telephone conversations using the ORD data, not for all fragments of recording it is possible to hear the voice of informant's interlocutor on the phone, unless the speaker changed the conversation mode to the speakerphone.
- Finally, real empirical data contain a certain share of fragments of speech, for which it is impossible to obtain an unambiguous text transcript or understand what is being said.

Nevertheless, despite the noted difficulties, manual tagging of speech data seems to be an important prerequisite for preparing test data for machine learning tasks, and the data obtained from its processing are good material for developing voice assistants, which behavior is as close as possible to real speech communication.

An important conclusion for the final part of the study is that the variety of replicas used by chat bots can lead to an undesirable result, since usually in everyday situations people use a very limited set of speech stimuli, the expansion of which makes them perceive communication as not quite familiar or natural.

As a practical recommendation for chat bot developers, we can suggest increasing the variability of answers not due to the lexical diversity of phrases, but due to a more diverse intonation implementation of the most typical replicas of spoken Russian.

ACKNOWLEDGMENT

The article was prepared with the financial support of the Grant of St. Petersburg State University (project No. 75254082 "Modeling the communicative behavior of residents of the Russian megalopolis in the socio-speech and pragmatic aspects with the involvement of artificial intelligence methods".

REFERENCES

[1] Google Assistant, Web: https://assistant.google.com/.
[2] Google Assistant Now Has 500 Million Users, Rivaling Amazon Alexa, Web: https://www.businessinsider.com/google-assistant-500-million-users-challenges-amazon-alexa-2020-1.
[3] Alice, Yandex voice assistant, Web: https://yandex.ru/alice.
[4] Siri, Apple, Web: https://www.apple.com/ru/siri/.
[5] Statista website. Number of voice assistants in use worldwide 2019-2024, Web: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.
[6] "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo, Web: https://www.researchgate.net/publication/316612010_Alexa_is_my_new_BFF_Social_Roles_User_Satisfaction_and_Personification_of_the_Amazon_Echo.
[7] A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova and T. Sherstinova, "The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation", *TSD 2009, Lecture Notes in Computer Science (LNCS)*, vol. 5729, 2009, pp. 250–257.
[8] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, O. Ermolova, E. Baeva, G. Martynenko and A. Ryko, "Sociolinguistic extension of the ORD corpus of Russian everyday speech", *SPECOM 2016, Lecture Notes in Artificial Intelligence (LNAI)*, vol. 9811. 2016, pp. 659–666.
[9] Demo-version of the ORD speech corpus, Web: https://ord.spbu.ru/.
[10] T. Sibata, "Studying the life of language with the method of the 24-hour survey" [Issledovanija jazykovogo sushhestvovanija v techenie 24 chasov], *Linguistics in Japan* [*Jazykoznanie v Japonii*], 1983, pp. 134–141.
[11] N. Campbell, "Speech & expression; the value of a longitudinal corpus", *LREC 2004*, pp. 183–186.
[12] L. Burnard, Reference guide for the British National Corpus, 2016. Web: http://www.natcorp.ox.ac.uk/docs/URG/.
[13] T. Sherstinova, "Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD Speech Corpus", *SPECOM 2015, Lecture Notes in Artificial Intelligence (LNAI),* vol. 9319, 2015, pp. 268–276.
[14] T. Sherstinova, "Approaches to pragmatic annotation in the ORD corpus: microepisodes and speech acts" [Pragmaticheskoe annotirovanie konnunicativnykh jedinic v korpuse ORD: mikroepisody i rechevye akty], *in Proc. "Corpus linguistics-2015" Conf.*, 2015, pp. 436–446.
[15] J.L. Austin, *How to do things with words*, Oxford: Oxford University Press, 1962.
[16] J.R. Searle, "A classification of illocutionary acts", *Lang. in Society*, vol. 5(1), 1976, pp. 1–23.
[17] W.B. Stiles, *Describing Talk: A Taxonomy of Verbal Response Modes*, Newbury Park: Sage Publications, 1992.
[18] J. Allen and M. Core, *Draft of DAMSL: Dialog act markup in several layers*, 1997. Web: https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.
[19] D. Jurafsky, "Pragmatics and computational linguistics", *The Handbook of Pragmatics*, 2006, pp. 578–604.
[20] H. Sloetjes and P. Wittenburg, "Annotation by category – ELAN and ISO DCR", *in Proc. LREC 2008 Conf.*, pp. 816-820.
[21] M. Weisser, "Speech act annotation", *Corpus Pragmatics: a Handbook*, 2014, pp. 84–111.
[22] I.N. Borisova, "Russian spoken dialogue. Structure and Dynamics" [Russkiy razgovornyy dialog. Structura i dinamika], *LIBROKOM*, 2009.
[23] T. Sherstinova, "Speech acts annotation of everyday conversations in the ORD corpus of spoken Russian", *SPECOM 2016, Lecture Notes in Artificial Intelligence (LNAI)*, vol. 9811, pp. 627–635.
[24] T. Sherstinova, "The structure of everyday dialogue as the sequence of speech acts" [Struktura povsednevnogo dialoga kak posledovatel'nost' rechevykh aktov], *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, Vol. 17, 2018, pp. 637-651.
[25] O.B. Ermolova and N.V. Bogdanova-Beglarian, "Linguistic design of "goodbye" situation in modern colloquial speech (based on the material of the speech corpus "one speaker's day")" [Yazykovoye

oformleniye proshchaniya sovremennoy razgovornoy rechi (na materiale zvukovogo korpusa “Odin Rechevoy Den’”)], *Communication Studies [Kommunikativnyye issledovaniya]*, vol. 6, iss. 2, 2019, pp. 307–331.

[26] V.N. Shaposhnikov, “Russian speech etiquette expressions. Phase formulas of modern communication”, *Journal of Philological Research [Zhurnal filologicheskikh issledovaniy]*, № 1, 2020, pp. 21–25.

[27] P. Lightbown, *How Languages are Learned*, Oxford: Oxford Handbooks for Language Teachers, 2012, p. 115.

[28] Russkiy yazyk povsednevnogo obshcheniya: osobennosti funktsionirovaniya v raznykh sotsial'nykh gruppakh [Russian language of everyday communication: features of functioning in different social groups, Web: http://www.ord-corpus.spbu.ru/SocialStudies/ORD.html.

[29] I.S. Nikolaev, O.V. Mitrenina and T.M. Lando, *Applied and Computational Linguistics [Prikladnaya i komp'yuternaya lingvistika]*, Moscow: Lenand, p. 320, 2006.