N-gram Analysis of Everyday Russian Speech: in Search of Multiword Units

Tatiana Sherstinova
[1]Saint Petersburg State University,
[2]National Research University Higher School of Economics,
Saint Petersburg, Saint Petersburg, Russia
tsherstinova@hse.ru

Olga Markovich
National Research University Higher School of Economics,
Saint Petersburg
Saint Petersburg, Russia
oimarkovich@edu.hse.ru

***Abstract*—Based on a statistical analysis of transcripts from everyday spoken Russian recordings, the presented research aims to search for stable multiword units. These units encompass a diverse set of multiword elements, bridging various linguistic phenomena such as compounds, idioms, colligations, collocations, collostructions, and multiword named entities. The n-gram analysis technique facilitates the identification of these units by capturing the most recurrent word sequences. Data for this research was sourced from the transcribed part of the ORD corpus, known as "One Speech Day", containing about 1,000,000 tokens. Captured using a continuous recording method with voluntary participants in natural conversational environments, this corpus is a best resource to study daily Russian dialogues. An examination of the top 500 bigrams and trigrams led to their categorization and the discernment of the most prevailing stable multiword units. These insights bear considerable relevance to NLP challenges centered on spontaneous Russian speech processing (primarily, for speech recognition tasks) as well as for teaching Russian as a second language.**

I. Introduction

Everyday spoken language is rich in idioms, speech clichés, and other *multiword units*. Traditionally, multiword units include units such as collocations — combinations of a word with other lexical elements, as well as colligations — combinations of words with specific grammatical markers or combinations of grammatical indicators of two or more words. These multiword units have a complex nature and perform various functions in speech (for instance, they can be used to make speech more vivid) [1].

For the Russian language, examples of these units include phrases like "*eto samoe*" (that very thing), "*bez problem*" (no problem), "*po barabanu*" (don't care), "*v tom-to i delo*" (that's the point), "*v samiy raz*" (just right), "*vykhodit' iz polozheniya*" (get out of a situation), "*otkuda nogi rastut*" (where it comes from), "*ne v sebe*" (not oneself), "*na khalyavu*" (for free), "*mama dorogaya*" (dear mother), "*Gospodi ty Bozhe moy*" (Oh, my God) and so on. For a range of tasks related to speech technologies (speech synthesis and recognition), machine translation, and studying Russian as a foreign language, it is crucial to have a relatively complete list of such units.

Empirical studies of spoken language show that multiword units (collocations, including idioms, colligations, collostructions, multiword named entities, etc.) are an integral part of everyday oral discourse, but they have not yet been systematically described based on Russian oral discourse material. Thus, for everyday Russian speech, there is no complete list of multiword units, despite the existence of a significant number of diverse resources and publications [2]; [3]; [4]; [5]; [6]; [7]; [8]; [9]; [10]; [16]. The reason for this is not only that the idiomatic expressions of oral texts differ from the idiomatic expressions of written texts, which most idiom dictionaries focus on. In the case of everyday spoken language, which is a living, evolving entity, the list of stable multiword units constantly changes, with new speech clichés and expressions emerging, such as "*Vse v shokolade*" (everything's great), "*Kak vse zapushcheno!*" (How everything is neglected!), "*Vypast' v osadok*" (be taken aback), "*vynos mozga*" (blow the brain), and others.

The aim of this research is to study the set of these complex linguistic phenomena in Russian everyday speech, which intersect lexicon and syntax, and also have a certain idiomaticity and statistical reproducibility [1]. Moreover, the study of these units is based on a data-driven approach, examining empirical material from contemporary speech recordings.

For the study of stable multiword units in our research, n-gram analysis is applied [11]. N-grams are sequences of text graphic units of the same level (most often letters or words), the frequency lists of which are commonly used in contemporary applied tasks of natural language processing. *N* can take any positive integer value and indicates the number of units considered in a sequence. For major NLP tasks, smaller values of *n* are most often used, ranging from 1 to 5. In our work, n-grams are used to count sequences of graphic words based on transcriptions of oral speech recordings. A graphic word is defined as any sequence of letters separated by a space or another non-letter character (e.g., a hyphen).

Most words in the Russian language can be categorized as 1-grams (unigrams), e.g., "*da*" (yes), "*privet*" (hello), "*utro*" (morning), "*doroga*" (road), "*zima*" (winter), and so on. Some words fall into the category of 2-grams, for instance, compound conjunctions: "*tak kak*" (because), "*potomu chto*" (because), "*kak budto*" (as if), and all hyphenated words: "*vitse-prezident*" (vice-president), "*mat'-geroinya*" (heroic mother), "*sekretar'-referent*" (secretary-assistant), "*shef-povar*" (chief-cooker), "*po-domashnemu*" (homemade), "*koe-chto*" (something), "*kogda-to*" (once), "*kak-nibud'*" (somehow), "*Sen-Sans*", "*Sankt-Peterburg*" (Saint Petersburg), etc. Some compound words are 3-grams — for instance, "*vslledstvie togo chto*" (due

to), "*Rostov-na-Donu*" (Rostov-on-Don). The lists of N-grams enable automatic identification not just of compound words but also concepts and named entities comprising multiple graphic words: "*Krasnaya ploshchad'*" (Red Square), "*Zimnyaya kanavka*" (Winter Canal), "*Bol'shoy dramaticheskiy teatr*" (Great Dramatic Theater), "*kursy povysheniya kvalifikatsii*" (professional development courses), etc. A simple frequency word dictionary (the list of unigrams) is insufficient for the automatic identification of such multiword units and idioms. So, n-gram analysis should be considered a handy tool to identify not only compound words and concepts but also frequent collocations, constructions, etc. [12].

In our research, we use n-gram analysis as a supplementary tool, employed to extract all possible sequences of graphic words based on a representative sample of oral speech transcriptions with the aim to get the list of the most frequent multiword units.

II. Studies and Classifications of Stable Multiword Units on Russian Empiric Data

This study continues a series of research on multiword units based on empirical audio material. The first significant study of this kind was the work of Dayang Liu [13].

In this research, the analysis of colloquial phraseology was conducted on the transcribed macroepisodes of everyday conversations of 20 informants from the ORD corpus [15] and their interlocutors. Informants were selected with 5 people in each gender and age group. Two age groups were identified as follows: older (≥ 40 years) and younger (< 40 years).

The research data consisted of 72 macro episodes of verbal communication with a total sound duration of about 22 hours and a total volume of text transcriptions of 230,000 words. This speech material was reviewed by experts, and all multiword elements that could be attributed to phraseological units were listed. The rather small sample size of this study requires considering the obtained statistics as preliminary, but it was the first research of this kind based on original Russian recordings. The results showed that the number of idioms in the total volume of speech material is not that large — in words, it constitutes only 0.29% and 0.28%, and the number of idioms per minute of recording is 0.48 and 0.52 for women's and men's speech, respectively.

A strong point of the work is the development of its own classification of multiword units. The following units were included:

1) **Codified idioms** are unquestionable: proverbs, sayings, idiomatic idioms – such as "*vopros na zasypku*" (a tricky or unexpected question), "*ne to slovo*" (I am quite agree with you), "*tyazhelyy sluchay*" (literal: heavy case), etc.

2) **Codified idiomatic exclamations**: "*Bog znaet chto!*" (That's terrible!), "*Da ty chto!*" (Are you serious?!), "*Gospodi ty Bozhe moy!*" (Oh my God!). These first two groups of idioms can be considered a kind of core of the idiomatic Russian spoken language.

3) **Idiom forms**: This is the third large group of stable multiword expressions from the core zone of Russian oral phraseology – e.g., "*ni o chem*" (of no value), "*za kompaniyu*" (just for company), "*do duri*" (to do smth to madness), etc.

4) **Modified idioms** are that which are not recorded in any dictionaries. These are non-codified (contextual) modifications and newly formed units that are potential idioms: "*bit' nogoy*" (literal: beat with the foot) instead of "*bit' kopytom*" (literal: beat with the hoof) meaning get angry or feel irritated. The author classifies this layer of non-codified material as the near periphery of the field structure of Russian colloquial phraseology.

5) **Idiomatized constructions** like "*S uma sosha chto li?*" (Have you lost your mind?), "*takoye vpechatleniye chto*" (it seems that).

6) **Occasional or contextual phraseological units**, and frequently used stable expressions that are not yet recorded in dictionaries formed not by modification of an existing idioms. For example, "*odnu sekundu*" (just a second), "*eto ya uzhe molchu*" (I'm not even talking about this), etc.

7) **Conversational variants of idiomatic interjections unregistered in dictionaries like** "*nu ty podumay!*" (Oh, my!).

8) **Precedent texts** in any language, which represent a whole unit and are also able to be replaced by an identifier unit — e.g., "*Alyo, garazh!*" (Hello, garage! — an allusion to a famous joke) and others like "*Ikh yest' u menia*" (I have them).

Modified idioms were included in the subcorpus with a special note when there was uncertainty about whether the speaker intentionally modified the idiom or if it was due to a speech error, i.e., characteristic of spontaneous speech production (this forms the distant periphery of the field phraseological structure of Russian spoken language).

Introductory constructions of various types (for instance, "*roughly speaking*", "*to put it another way*", etc.) were not included as idioms/FE (in the user subcorpus). This is because they, like pragmatic markers [14]; [17], are not actual speech units but conditionally-speech functional units of oral discourse.

Units for analysis were searched for using a method of continuous sampling, reviewing the speech material with a record of the duration of the sound and the volume in words of each speech fragment. This provided a preliminary list of such units, statistics on their usage in speech, and also allowed for examining the dependence of the appearance of idiomatic/phraseological units of different types on speaker characteristics and on the communicative situation as a whole.

This study was entirely expert-based and provided preliminary classification and statistics of stable multiword units. However, it was followed by another study that used N-gram analysis for typology construction [18].

The source material for applying this methodology was a selection of 388 episodes of everyday verbal communication from the same ORD corpus, but on other speech data (about 110 hours of audio). Based on this material, two frequency lists were obtained – bigrams and trigrams (see Table I-II). The typology of the most commonly used bigrams and trigrams in Russian spoken communication was found based on the top 200 units from these frequency lists. These are as follows:

1) **Vocalizations (VOK)** ("*e-e*", "*m-m*", "*a-a*") — a variety of hesitation phenomena, one of the ways of non-verbal filling of hesitation pauses. Vocalizations are understood as "speech-

like" sounds, or sounds of a "non-phonemic nature". Such elements are typically considered a form of speech disruption, where the smooth flow of speech is interrupted. This type of disruption is a "break used by the speaker to prepare for the next portion and/or (in combination with correction) to consider a possible way to correct the previous portion" (Podlesskaya, Kibrik 2005).

2) **Amplifications (AMPL)** ("*da-da*") - specific repetitive units, often formed by syllable repetition, such as "*op-op-op*", "*to-to-to*", "*ta-ta-ta*", "*tak-tak-tak*", etc. In everyday speech, they often act as pragmatic markers (for more details on the class of amplified units, see [19]).

3) **Compound conjunctions (CONK)** ("*to est'*", "*potomu chto*").

4) **Pragmatic markers (PM)** (basic multiword units, their structural variants, or chains) ("*ne znayu*", "*nu vot*", "*kak by*") (for more details on PMs, see the specialized dictionary of such units [17]).

5) **Single-word lemmas (LEMMA)** ("*kak-to*", "*chto-to*").

6) **Combinations of two particles (2 PART)** ("*nu da*", "*vot eto*").

7) **Bigrams**: in terms of these classification, combinations of particles with other words, not linked by any relations **(BIGRAM)** ("*ya ne*", "*nu ya*").

8) **Actual grammatical structures:** Prepositional-case word forms **(PPF)** ("*u menya*", "*u nas*", "*u tebya*"), Predicative base **(PREDIC)** ("ya govoryu").

The study showed that the most frequent 20 bigrams and trigrams in the oral speech sample were the following (see Table I-II) [18]:

TABLE I. MOST FREQUENT BIGRAMS OF EVERYDAY RUSSIAN SPEECH (TOP-20)

Rank	Type	Freq	NormFreq = ipm	Status
1	*e-e*	3746	4217	VOK
2	*u menya*	2157	2428	PPF
3	*to yest'*	1926	2168	CONC
4	*u nas*	1635	1840	PPF
5	*ya ne*	1572	1769	BIGRAM
6	*potomu chto*	1551	1746	CONC
7	*da-da*	1525	1717	AMPL
8	*ne znayu*	1512	1702	PM
9	*nu vot*	1338	1506	PM
10	*m-m*	1312	1477	VOK
11	*kak by*	1252	1409	PM
12	*chto-to*	1166	1312	LEMMA
13	*u tebya*	878	988	PPF
14	*nu da*	866	975	2 PART
15	*vot eto*	856	964	PM
16	*a-a*	834	939	VOK
17	*kak-to*	779	877	LEMMA
18	*nu ya*	775	872	BIGRAM
19	*vot tak*	761	857	PM
20	*ya govoryu*	737	830	PREDIC

This study showed that the n-gram analysis method employed has proven effective in providing raw data for classifying bigrams and trigrams. It allows to shed light on grammatical patterns, fixed expressions in verbal communication, and pragmatic markers. Moreover, it suggests various other ways to analyze corpus data. This method precisely depicts the balance between lexicogrammatical and pragmatic elements in speech and the distinction between major and minor components. Understanding this balance is vital for an in-depth study of oral communication.

TABLE II. MOST FREQUENT TRIGRAMS OF EVERYDAY RUSSIAN SPEECH (TOP-20)

Rank	Type	Freq	NormFreq = ipm	Status
1	da-da-da	688	774	AMPL
2	ya ne znayu	557	627	PM
3	na samom dele	286	322	PM
4	vot tak vot	264	297	PM
5	nu v obshchem	210	236	PM
6	vot e-e	190	214	PM + VOK
7	ya dumayu chto	170	191	PM
8	vot eto vot	167	188	PM
9	potomu chto ya	144	162	TRIGRAM
10	nu kak by	139	156	PM
11	e-e nu	138	155	VOK + PART
12	e-e-e	138	155	VOK
13	i daleye	135	152	PM
14	nu ne znayu	126	142	PM
15	e-e v	124	140	VOK + PROPOSITION
16	nu to yest'	123	138	TRIGRAM
17	a u menya	114	128	TRIGRAM
18	nu i chto	113	127	IDIOM
19	chto u nas	112	126	TRIGRAM
20	m-m-m	105	118	VOK

Through this methodology, a refreshed view of oral discourse's grammar emerged. It becomes evident that not just the notable grammatical constructs but also the common sequences of words play a pivotal role. The disparity between these significant and less significant elements, crucial from a pragmatic standpoint, is so vast that sidelining the latter would be a misstep. Both automated speech recognition systems and individuals unfamiliar with the Russian language process the complete auditory sequence of the conversation, not merely its meaningful segments. Grasping a conversation commences with discerning this overall auditory content, which hinges not only on distinguishing the primary from the secondary but also on recognizing common auditory patterns. Recognizing these patterns can assist in efficiently navigating the auditory flow of speech [18].

This research continues the previous studies, being conducted on more representative material and offering additional variants for n-gram categories. The research is primarily interested in the semantic features of n-grams, their role in speech, which is reflected in the creation of a preliminary classification that could be further detailed in the future.

III. DATA AND METHOD

The calculations were based on a sample of 463 episodes of everyday spoken communication from the ORD corpus with recordings made in 2007 and from 2014 to 2016 [20]; [21]; [22]. The total speech duration, excluding extended pauses, amounts to about 250 hours. Speech transcripts contains about 800,000 tokens. Selected episodes capture the full spectrum of daily spoken communication in Russian – everyday household

conversations, professional communication at work, informal chats with colleagues, interactions with friends, acquaintances, and relatives, as well as varied verbal exchanges in customer-service settings like shops, medical centers, customer service departments, etc. [23]. The recordings come from informants of various social and professional background [24].

Transcriptions for the ORD corpus were made manually using the ELAN multimedia annotation environment [25] and are stored in its format (*.eaf). For automatic extraction and counting of n-grams, the "Phrases" level was extracted from the transcriptions [21]. The Phrase level contains major information recorded by the microphone, i.e., human speech, various pauses, as well as other paralinguistic sounds (laughter, coughing, yawns). The transcriptions underwent preprocessing while retaining information about phrase boundaries, lines, and speaker shifts in overlapping speech segments.

IV. The Most Frequent N-Grams

The calculations were based in AntConc corpus manager [26]. In Table III, the top zones of the most frequent n-grams are presented (for n=2 and 3).

The obtained frequency lists demonstrate a significant overlap with prior results from everyday spoken language. This suggests a hypothesis that within a specific genre, n-gram frequency lists exhibit a high degree of consistency [27]. Amplificated phatic elements, such as "*da-da*" (*yes-yes*), "*da-da-da*" (*yes-yes-yes*), and "*ugu-ugu*", as well as hesitations and compound words "*to yest'*" (*that is*), "*potomu chto*" (*because*), dominate the top of this list. We can also note such units as "*ne znayu*" (*I don't know*), "*na samom dele*" (*actually*), "*a chto*" (*so what*), "*vot tak vot*" (*just like that*). The obtained lists indicate that the share of the multiword units of interest to us at the top of the frequency list is not large, with a greater prevalence observed in trigrams compared to bigrams.

To navigate the wide variety of n-grams, we propose a new classification system based on their pragmatic function. Employing this system for expert manual annotation will facilitate the creation of a training dataset. This will enable the automated extraction of multiword units from extensive text transcriptions, thus advancing our outlined objectives.

V. Developing a Classification Framework for N-Gram Annotation

N-grams are automatically generated units, and they often lack semantic unity, which poses challenges for annotation. Hence, the nature of an n-gram will be determined by the semantics of its main word, all senseless combinations will be categorized distinctly. The development of n-gram classification will lean on the typology of pragmatic markers proposed by N. Bogdanova-Beglaryan [28].

A salient observation is that the majority of the identified n-grams play a structuring role in speech, acting as contextual markers. These derived lists predominantly consist of speech units consistently present in dialogues. They facilitate message delivery, structure its narrative, pinpoint participants, outline their actions, attribute thoughts, and convey the speaker's opinion toward an event or individual.

Upon initial inspection of the acquired n-grams, it's evident that live spontaneous speech is replete with hesitative elements.

TABLE III. Most Frequent N-Grams

Rank	N-grams	Translation	Norm. frequency (ipm)
2-grams			
1	*eh eh*	*uh huh*	5878.986
2	*da da*	*yes yes*	3249.184
3	*u menya*	*I have*	2995.020
4	*to yest'*	*that is*	2648.120
5	*u nas*	*we have*	2396.245
6	*ya ne*	*I don't*	2263.438
7	*potomu chto*	*because*	2245.120
8	*ne znayu*	*I don't know*	2119.183
9	*nu vot*	*well*	1977.217
10	*mm*	*uh huh*	1822.657
11	*kak by*	*kind of*	1791.745
12	*chto to*	*something*	1759.689
13	*a a*	*uh*	1480.337
14	*nu da*	*yes*	1419.658
15	*vot eto*	*this*	1352.109
16	*vot tak*	*like this*	1329.212
17	*da nu*	*really*	1310.894
18	*u tebya*	*you have*	1242.200
19	*nu ya*	*well, I*	1233.041
20	*a ya*	*and I*	1158.624
21	*a chto*	*so what*	1104.814
22	*kak to*	*somehow*	1088.786
23	*v obshchem*	*generally*	1071.613
24	*ya govoryu*	*I'm saying*	1055.584
25	*chto ya*	*that I*	1051.005
3-grams			
1	*da da da*	*yes yes yes*	1215.870
2	*ya ne znayu*	*I don't know*	788.827
3	*vot tak vot*	*like this*	457.955
4	*na samom dele*	*actually*	398.421
5	*vot eh eh*	*well, uh, uh*	361.784
6	*nu v obshchem*	*well, in general*	311.409
7	*eh eh eh*	*uh, uh, uh*	289.656
8	*eh eh nu*	*uh, uh, well*	285.077
9	*vot eto vot*	*this here*	277.063
10	*eh eh v*	*uh, uh in*	248.440
11	*ya dumayu chto*	*I think that*	231.267
12	*nu kak by*	*well, sort of*	218.673
13	*potomu chto ya*	*because I*	212.949
14	*i tak dalee*	*and so on*	204.935
15	*da to est'*	*yes, that is*	201.500
16	*a u menya*	*I have*	191.196
17	*nu ne znayu*	*well, I don't know*	188.906
18	*eh eh vot*	*uh, uh, look*	177.457
19	*nu to est'*	*well, that means*	174.023
20	*chto u nas*	*what we have*	174.023
21	*da da nu*	*yes, yes, well*	170.588
22	*m m m*	*mmm mmm mmm*	169.443
23	*ne znayu ya*	*I don't know*	168.298
24	*nu i chto*	*well, what about it*	167.153
25	*ugu ugu ugu*	*uh-huh uh-huh uh-huh*	167.153

This means speakers frequently employ certain lexemes, clitics, and non-verbal elements, likely to provide a pause for thought. A holistic view of bigrams and trigrams reveals various forms of hesitation in spontaneous speech: linguistic; emotional; cognitive.

When addressing linguistic hesitation, a speaker introduces a verbal pause due to uncertainty about the correct word choice or lexical arrangement. In contrast, emotional hesitation manifests when a speaker is ambivalent about their feelings in relation to a situation. Cognitive hesitation, meanwhile, emerges when there's uncertainty in the precision and authenticity of one's thoughts and convictions. It's imperative to

acknowledge that in contemporary spoken language, some hesitative expressions have evolved into standard markers that either commence or conclude a statement. For instance, the introductory "well" or the concluding "you know" act as rhythmic anchors, leading to potential homonymy issues.

In the realm of European discourse, linguistic hedging—a strategy where interlocutors regularly employ precautionary linguistic tools in spontaneous expression—is pervasive [29]. This includes the use of modal verbs, judgmental adverbs, double negatives, and indefinite pronouns and adverbs. Such devices often serve to diminish the speaker's liability, indicating their intent to circumvent unequivocal information and underscore the subjectiveness of their remarks. These lexical units, due to their role in indicating uncertainty, can be categorized under hesitations.

Analyzing the n-gram list, it's evident that 30% of the most recurrent bigrams and 48% of trigrams have elements of hesitation or rhythmic constituents. These two categories often overlap since a single linguistic unit can fulfill dual roles, or an n-gram can encompass elements from both categories.

Assigning classification tags to n-grams is also challenging, primarily due to the inherent ambiguity or homonymy of many n-grams. Absolute precision is unattainable, so it's essential to clarify that the categories designated during the analysis reflect the most frequent meanings. Another issue, previously mentioned, concerns the data collection method for research and the ensuing semantic discrepancies. These instances can be classified into several types:

The first group comprises n-grams that form part of stable combinations. For example, in "*na samom, samom dele*" (literally "*in the very, very fact*"), which is a part of the idiomatic phrase "*na samom dele*" (*actually* or *in fact*). The incompleteness of the unit can be explicit, as in the given example, or implicit, becoming apparent only upon examining the context.

The second group entails units that are difficult to identify: n-grams positioned at phrase boundaries that don't form a semantic whole.

Lastly, the third group includes semantically incomplete speech elements. Common examples are bigrams that contain a preposition without its corresponding dependent word, like "*i na*" (*and on*) or "*nu v*" (*well in*), or the particle "*by*". Determining the function of the preposition in such n-grams is especially challenging since prepositions possess a high combinatory potential and can form semantically diverse phrases. For instance, they might indicate time or refer to an object, as seen in the difference between "*na samom dele*" (*actually*), "*na dnyakh*" (*in the coming days*), and "*na eto*" (*for this* or *on this*).

In addition to idioms, it's worth paying attention to the study of constructions, that is, combinations of units that can be considered stable [30]. During the n-gram analysis, several such combinations were found:

Bigrams: "*v printsipe*" (*in principle*), "*to yest'*" (*that is*), "*do svidaniya*" (*goodbye*), "*v smysle*" (*in the sense*), "*v obschem*" (*in general*), "*mozhet byt'*" (*maybe*), "*chut'-chut'*" (*a little bit*), "*vo-pervykh*" (*firstly*);

Trigrams: "*na samom dele*" (*actually*), "*po krayney mere*" (*at least*), "*v lyubom sluchaye*" (*in any case*), "*na vsyakiy sluchay*" (*just in case*).

Furthermore, it's interesting to examine the position of trigrams within a phrase specifically since they offer more detailed insights, and the results are more indicative. As expected, most of the trigrams (58%) are located at the beginning of a statement. This can be linked to the function of the identified n-grams. This reflects the typical structuring of speech: due to the spontaneity of utterances, people require familiar phrases to initiate the speech production process.

The detailed analysis of empiric data obtained led to the following classification scheme:

1. Discourse markers,
2. Phatic markers,
3. Metacommunicative elements,
4. Hesitations,
5. Referentials,
6. Subordinators,
7. Start/end markers,
8. Relation markers,
9. Semantically incomplete combinations, predominantly with prepositions and conjunctions.

A brief description of each category is given in the next section.

VI. Statistical Analysis Results: Frequencies for Specific N-Gram Classes

This section presents statistics for each of the identified categories of multiword units, supplied with information on the actual frequency statistics for the top 500 bigrams and trigrams.

1) **Discourse markers** are understood as units that structure speech [31]. Notably, most of all the previously mentioned expressions belong to this category. This is likely because the other classes consist of more flexible units that do not suggest idiomaticity. The overall percentage for the relative frequencies of bigrams and trigrams is 5.43% and 11.71% respectively.

Table IV provides an example of the top zone of discourse markers. Similar statistics were obtained for each of the categories.

2) **Phatic markers** are units that imply a preceding or following statement from the interlocutor. They include units that express affirmation, negation, and interrogatives, and they are most often found at the beginning of phrases. Examples include ("*ugu ugu ugu*" which means "*uh-huh uh-huh uh-huh*", "*da da ya*" which means "*yes yes I*", "*net u menya*" which means "*I have no*", "*chto eto takoe*" which means "*what is this*"). Their percentage of relative frequency in speech is 17.81% for bigrams and 20.41% for trigrams.

3) **Metacommunicative elements**. This category encompasses a wide range of n-grams: firstly, those containing metacommunicative verbs (like "*say*", "*think*", "*know*", etc.). Secondly, it includes participants of communication at various

TABLE IV. MOST FREQUENT DISCOURSE MARKERS

Rank	N-grams	Translation	Norm. frequency (ipm)
2-grams			
1	*to est'*	*that is*	2648.12
2	*v obshchem*	*in general*	1071.613
3	*tak vot*	*so*	918.198
4	*vsyo ravno*	*anyway*	563.284
5	*v printsipe*	*in principle*	507.184
6	*vsyo taki*	*after all*	428.187
7	*na samom*	*actually*	414.448
8	*samom dele*	*in fact*	409.869
9	*a tak*	*otherwise*	291.946
10	*v smysle*	*in terms of*	281.642
11	*vo pervykh*	*firstly*	196.92
12	*znachit eh*	*so, uh*	190.051
13	*esli chto*	*if anything*	149.98
14	*do svidaniya*	*goodbye*	139.676
3-grams			
1	*na samom dele*	*actually/in fact*	398.421
2	*nu v obshchem*	*well, generally*	311.409
3	*da to yest'*	*yes, that is*	201.500
4	*nu to yest'*	*well, that is*	174.023
5	*v obshchem to*	*in general*	122.503
6	*znachit eh eh*	*means*	119.068
7	*to yest' eh*	*that is*	89.301
8	*to yest' on*	*that is he*	85.866
9	*po krayney mere*	*at least*	74.418
10	*to yest' u*	*that is it*	74.418
11	*s drugoy storony*	*on the other hand*	73.273
12	*ugu to yest'*	*uh-huh, that is*	70.983
13	*delo v tom*	*the thing is*	68.693
14	*nu v printsipe*	*well, in principle*	67.548
15	*v lyubom sluchaye*	*in any case*	65.259
16	*i v obshchem*	*and generally*	60.679
17	*imeyu v vidu*	*I mean*	60.679
18	*tem ne menee*	*nevertheless*	60.679

levels: metacommunicative subjects (like "*I*", "*you*", "*we*") and narrative ones (like "*he*", "*she*", "*it*", "*they*"), as well as categories of belonging, divided in the same way. The percentage for bigrams is 29.67%, and for trigrams, it's 29.88%.

4) **Hesitations** are markers of uncertainty. Surprisingly, the frequency percentages of pure hesitations are almost identical to the previous categories: 9.72% for bigrams and 13.19% for trigrams.

5) **Referentials** contain n-grams that refer to time, or point to something in the real world: an object, an occurred situation, or a place. There is a high likelihood of homonymy with hesitatives or rhythmic markers, as it's hard to determine whether the person is genuinely referencing something or formulating a thought, filling pauses. The relative frequency percentages of this group are 14.18% for bigrams and 9.11% for trigrams.

6) **Subordinating conjunctions** include conjunctions of complex sentences (like "*chto*" (*that*), "*esli*" (*if*), "*khotya*" (*although*), "*potomu chto*" (*because*). Here, bigrams have a relative frequency of 3.55%, and trigrams have 5.69%.

7) **Start and End Markers.** This separate category consists of units that, in meaning, often lie between hesitations and discourse markers. Their relative frequency percentages are 3.96% for bigrams and 1.46% for trigrams.

8) **Relationship Markers.** This category of n-grams can be considered a subclass of discourse markers. They express various emotions of the speakers. The relative frequency for these n-grams are 1.74% for bigrams and 0.99% for trigrams.

9) Lastly, **semantically incomplete combinations**, predominantly with prepositions and conjunctions. In this class of n-grams, there are bigrams and trigrams with a relative frequency of 5.79% and 3.39% respectively, which can be considered functionally incomplete since they lack additional information for completeness.

VII. CONCLUSION

N-grams of spoken language are used in a wide range of tasks related to speech analysis, recognition, and generation. The presented research utilizes them to identify the most frequent stable multiword units in contemporary Russian spoken language, based on a representative sample of audio recordings. An expert examination of the top 500 bigrams and trigrams led to their categorization and the construction of a classification scheme, dividing these units into semantic-pragmatic groups. The proposed scheme comprises 9 main categories.

The study revealed that many idioms and idiom-like expressions, frequently found in dictionaries of "stable and winged expressions" and described earlier in section II, are missing from the top zone of the frequency dictionary. They are present in spoken language but occur significantly less frequently than discursive and phatic markers, hesitations, markers of beginnings and ends, and other frequent elements of spoken speech. The share of stable expressions among the list of N-grams turned out to be small, as shown by the obtained statistics.

Therefore, the proposed classification scheme should be considered an expansion of the traditional understanding of multiword units categories, described in work [13]. As for the pilot classification proposed in work [18], this 9-category scheme relates exclusively to multiword units, therefore better aligns with the goals set before our research.

The significance of the conducted study lies not only in identifying the most frequent multiword units and their categories but also in developing an annotation scheme, which is planned to be expanded to a larger volume of speech material in the future. Functional categories with tagging can be used for machine learning to detect less frequent multiword units.

Continuing the research could involve identifying the most frequent components within n-grams. Thus, Gris, reflecting on half a century of work on collocations, emphasized the importance of distinguishing between combinations of words with strong and weak ties [32]. This suggests that certain words will inherently form part of many n-grams, a notion supported by our findings. As a continuation of this research, it seems reasonable also to study the typical ways of forming multiword units, which would allow recognizing these units even if they are absent in the dictionary or training sample, and to predict their occurrence. Another important aspect is comparing the function of multiword units and its position in the linear unfolding of the phrase. It can be hypothesized that certain

categories of multiword units will tend to have a specific position within the phrase, which could also serve as a marker for their identification.

Employing the devised classification during expert manual annotation can facilitate the creation of a dataset for automatic extraction of stable multiword units from extensive text transcriptions. This approach inches closer to the pivotal practical objective of pinpointing a comprehensive list of stable multiword units. Such steps are particularly pertinent to NLP tasks related to spontaneous Russian speech processing, especially for speech recognition and for instructional methodologies in teaching Russian as a foreign language.

ACKNOWLEDGMENT

The presented research was supported by the Russian Science Foundation, project No. 22-18-00189 “Structure and functionality of stable multiword units in Russian everyday speech”.

REFERENCES

[1] Bogdanova-Beglarian N. V., Blinova O. V., Khokhlova M. V., Sherstinova T. Yu. Towards the Description of Multiword Units in Russian Everyday Speech: State-of-the-Art and the Methodology of Further Research. In: Mukhamediev R., Pereira R., Mityagin S., Bolgov R. (eds.) Springer Geography, Springer Nature, 2023. Pp. 129-139.

[2] Leksikograf, http://lexicograph.ruslang.ru/02prjtheorytext.htm, last accessed 2022/04/15.

[3] Lyashevskaya, O., Kashkin, E.: FrameBank: a database of Russian lexical constructions. In: M.Yu. Khachay, N. Konstantinova, A. Panchenko, D.I. Ignatov, G.V. Labunets (eds.), Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015, Revised Selected Papers. Communications in Computer and Information Science, Vol. 542, Springer, pp. 337-348 (2015).

[4] CoCoCo – “Collocations, Colligations, Corpora”, https://cosyco.ru/cococo/, last accessed 2022/04/15.

[5] Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., Yangarber, R. 'CoCoCo: Online Extraction of Russian Multiword Expressions'. The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, pp. 43-45 (2015).

[6] Biriuk, O.L., Gusev, V.Ju., Kalinina, E.Ju.: Dictionary of Russian Abstract Nouns’ Verbal Collocability. A Dictionary based on the Russian National Corpus.(In Rus.) = Slovar’ Glal’noj Sochetaemosti Nepredmetnykh Imen Russkogo Jazyka. Slovar’ na osnove Natsional’nogo Korpusa Russkogo Jazyka. Available at http://dict.ruslang.ru (2008).

[7] Kustova, G.I.: Dictionary of Russian Idiomatic Expressions.(In Rus.) = Slovar’ russkojidiomatiki. Sochetanijaslov so znachenijemvysokojstepeni. Available at http://dict.ruslang.ru (2008).

[8] Apresyan, Yu.D. (ed.): Aktivnyy slovar' russkogo yazyka [The Active Dictionary of the Russian Language] Vol. 1. A—B, Moscow, Yazyki slavyanskoy kul'tury [Languages of Slavic culture] (2014).

[9] Radovan, B., Endresen, A., Janda, L., Lund, M., Lyashevskaya, O., Mordashova, D., Nesset, T., Rakhilina, E., Tyers, F., Zhukova, V.: The Russian Constructicon. An electronic database of the Russian grammatical constructions. Available at https://constructicon.github.io/russian/Constructicon (2021).

[10] Endresen, A., Zhukova, V., Mordashova, D., Rakhilina, E., Lyashevskaya, O.: Russkiy konstruktikon: Novyy lingvisticheskiy resurs, yego ustroystvo i spetsifika [The Russian Constructicon: A new linguistic resource, its design and key characteristics]. In: Computational linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue-2020”. Issue 19, pp. 226-241 (2020).

[11] Jurafsky, D., Martin, J. H. (2009). Speech and language processing. London [u.a.]: Prentice Hall, Pearson Education International.

[12] Sherstinova, T.Yu. (2019). The Syntax of Everyday Russian Speech through the Prism of N-gram Analysis. In: Russian Grammar: Structural Organization of Language and Processes of Language Functioning. Ed. O.I. Glazunova, K.A. Rogova. Moscow: LENAND. Pp. 454-466.

[13] Liu, Dayang (2019). Phraseological Units in the Russian Everyday Speech: Typology and Functioning. PhD Thesis. St. Petersburg. 389 p. (typescript).

[14] Bogdanova-Beglarian, N.V., Blinova, O.V., Troshchenkova, E.V., Sherstinova, T.Yu., Gorbunova, D.A., Zaides, K.D., Popova, T.I., Sulimova, T.S. (2021). Pragmatic Markers of Russian Everyday Speech: Quantitative Data. In: Computational Linguistics and Intelligent Technologies. Issue 20 (27). Based on the Materials of the Annual International Conference "Dialogue" (2021). Ch. ed. V.P. Selegey. Moscow: RGGU. Pp. 119-126.

[15] ORD corpus of Russian everyday speech [Electronic resource], https://ord.spbu.ru/, last accessed 2023/04/01.

[16] National Corpus of the Russian Language [Electronic resource], https://ruscorpora.ru/, last accessed 2023/04/01.

[17] PM – Pragmatic Markers of Russian Everyday Speech: Dictionary-Monograph (2021). Ed. N.V. Bogdanova-Beglarian. St. Petersburg: Nestor-History. 520 p.

[18] Khokhlova M., Blinova O., Bogdanova-Beglarian N., and Sherstinova T. On the most frequent sequences of words in Russian spoken everyday language (bigrams and trigrams): an experience of classification. In: SPECOM 2023, LNCS, 14338/14339, Springer, 2023, Pp. 456-468.

[19] Sherstinova, T.Yu. (2016a). On Repetitions of Discursive Words in Everyday Speech Communication (Based on the Russian Language). In: Proceedings of the 45th International Philological Conference (IPC-2016). Volume 122. Ser. Advances in Social Science, Education and Humanities Research (ASSEHR). Eds. S. Monakhov, I. Vasilyeva, M. Khokhlova. Atlantis Press. Pp. 480-483.

[20] 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, pp. 250–257 (2009).

[21] 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, pp. 258–265 (2009).

[22] 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, pp. 659–666 (2016).

[23] 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. Springer, Switzerland, 2015, pp. 268–276

[24] Russian Language of Everyday Communication: Features of Functioning in Different Social Groups (2016). Collective Monograph. Ed. N.V. Bogdanova-BeglarIan. St. Petersburg: LAIKA. 244 p.

[25] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006 Fifth International Conference on Language Resources and Evaluation. Genoa, Italy. Pp. 1556-1559.

[26] Anthony, L. (2023). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University, http://www.laurenceanthony.net/software.

[27] Martynenko G. Osnovy stilemetrii (Foundations of stilometrics). Leningrad, 1988.

[28] Bogdanova-Beglaryan N. V. Pragmatemy v ustnoy povsednevnoy rechi: opredelenie ponyatiya i obshchaya tipologiya //Vestnik Permskogo universiteta. Rossiyskaya i zarubezhnaya filologiya. 2014. №. 3. S. 7-20.

[29] Gorina O. G., Khrabrova V. E. Lingvisticheskiy khedzhing kak kommunikativnaya strategiya (v rusle korpusnykh issledovaniy) // Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya. 2017. №3. S. 44–53.

[30] Rakhilina E. V. Lingvistika konstruktsiy / Otv. red. E. V. Rakhilina. M.: Izdatelskiy tsentr "Azbukovnik", 2010.
[31] Bogdanova-Beglarian N. V., Filyasova Yu. A. Discourse vs Pragmatic Markers: A Contrastive Terminological Study // 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts, SGEM 2018. Vienna ART Conference Proceedings, 19—21 March, 2018. Vol. 5, Iss. 3.1. P. 123—130.
[32] Gries St. 50-something years of work on collocations: What is or should be next // International Journal of Corpus Linguistics. 2013. Vol. 18. P. 137.